FULLY-AUTOMATED QUANTITATIVE ANALYSIS OF CARDIAC AND LUNG DISEASES FROM THORACIC LOW-DOSE CT IMAGES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Yiting Xie

August 2017

© 2017 Yiting Xie

ALL RIGHTS RESERVED

FULLY-AUTOMATED QUANTITATIVE ANALYSIS OF CARDIAC AND LUNG DISEASES FROM THORACIC LOW-DOSE CT IMAGES

Yiting Xie, Ph.D.

Cornell University 2017

Quantitative image biomarkers are emerging as a method for precision medical diagnosis. Fully-automated computer algorithms are explored to provide clinically useful biomarker measurements for the assessment of cardiovascular and lung diseases from low-dose thoracic computed tomography (CT) images. The recent regulatory approval of annual lung cancer screening (LCS) provides the opportunity for the application of these methods to a large at-risk population that will already be receiving annual low-dose chest CT scans. These computer algorithms must specifically address the high image noise levels concordant with the low-dose imaging protocol.

Quantitative evaluation of cardiovascular disease is facilitated by automated segmentations of cardiac organs (aorta, heart region, pulmonary trunk); primarily coronary artery calcification (CAC), a major indicator of coronary heart diseases, is scored. For lung disease assessment, the automated detection of interstitial lung disease (ILD) at its earliest detectable stage is performed. In addition, CT image quality (noise and calibration) is automatically assessed from segmented homogeneous regions for quality control and increased measurement precision.

Automated CAC scores have shown a 0.90 correlation with reference measurements provided by radiologists. The automated ILD detection algorithm is able to distinguish between early-stage ILD and normal cases with an Area Under the ROC curve of 0.95. The image quality assessment method has also shown to be repeatable and robust when evaluated on phantom images and a large LCS cohort.

This research advances the state-of-the-art of computer algorithms for precise region

segmentation and biomarker measurements that permit the evaluation of cardiac and lung health in the context of LDCT. The successful outcomes of these algorithms have demonstrated the possibility of automated chest health monitoring on an annual basis for a large population through the LCS process.

BIOGRAPHICAL SKETCH

Yiting Xie joined the Cornell University Vision and Image Analysis Group (VIA group) headed by Dr. Anthony Reeves in 2012. Her research work in the VIA group focuses on the fully-automated assessment of cardiac and lung biomarkers from low-dose chest CT images. Before coming to Cornell University, she received a Master's of Science degree in Electrical and Computer Engineering from Carnegie Mellon University in 2012, during which she worked in the Cylab Biometrics Center headed by Dr. Marios Savvides. She received her Bachelor's of Science degree in Electrical Engineering from Huazhong University of Science and Technology in Wuhan, China in 2010.

ACKNOWLEDGMENTS

I would like to thank my academic advisor, Dr. Anthony Reeves, for his guidance and support. Through his advice, I have gained some of the most valuable knowledge on how to think constructively and write succinctly. I am grateful for all his help and suggestions. Without him, I would not have been able to achieve this.

I would also like to express my gratitude to my colleague and friend Shuang Liu. Our discussions and exchanges of ideas have greatly helped me in my research. I truly appreciate her creativity and her unique insights.

I am grateful to my committee members Drs Tsuhan Chen and Noah Snavely for their time and encouragement; Drs. David Yankelevitz and Claudia Henschke at the Mount Sinai Hospital for their valuable feedbacks and the resources they have provided. I am fortunate to have worked with many colleagues from the Mount Sinai Hospital: Drs. Artit Jirapatnakul, Yu Maw Htwe and Mingzhu Liang. I am impressed by their work ethics and am grateful for their help. I would also like to thank the Flight Attendants' Medical Research Institute and the Cornell Neurotech Mong Family Foundation for their generous support.

Finally, I would like to thank my grandmother Yulian and my grandfather Xuancheng, for always being there for me; my friends Deepa, Jennifer, Feven, and Joya, for bringing so much joy to my life; and my professor Dr. Marios Savvides at Carnegie Mellon University, for introducing me to computer vision.

TABLE OF CONTENTS

Biographical Sketch ..... **iii**

Acknowledgments ..... **iv**

Table of Contents ..... **x**

List of Figures ..... **vii**

List of Tables ..... **x**

1. Introduction **1**
- 1.1 Low-dose chest CT from lung cancer screening 2
  - 1.1.1 Lung cancer and lung cancer screening 3
  - 1.1.2 Characteristics of low-dose CT 5
- 1.2 Computer aided diagnosis in low-dose CT: literature review and challenges 6
  - 1.2.1 CAD in the cardiac region 8
  - 1.2.2 CAD in the lung region 10
  - 1.2.3 CT image quality assessment 15
- 1.3 Segmentation using the Anatomy Label Map 16
- 1.4 Overview 19

2. Cardiac segmentation and disease biomarker measurement **20**
- 2.1 Segmentation of the aorta, the pulmonary artery trunk and the heart region 21
  - 2.1.1 Segmentation of the aorta 21
  - 2.1.2 Segmentation of the pulmonary artery trunk 25
  - 2.1.3 Segmentation of the heart region 27
- 2.2 Segmentation and measurement of cardiac biomarkers 28
  - 2.2.1 Segmentation and measurement of CAC 28
  - 2.2.2 Measurement of aorta diameter profile and PT diameter 33
  - 2.2.3 Segmentation and measurement of AC 34
  - 2.2.4 Segmentation and measurement of cardiac visceral fat 36
- 2.3 Quantitative and qualitative evaluation methods 38
  - 2.3.1 Quantitative evaluation experiments 38
  - 2.3.2 Qualitative evaluation experiments 43
- 2.4 Evaluation results 46
  - 2.4.1 Quantitative evaluation results 46
  - 2.4.2 Qualitative evaluation results 52
- 2.5 Discussion 52
- 2.6 Conclusion 55

3. Lung disease biomarker measurement **57**
- 3.1 Pulmonary nodule characterization 57

3.1.1 Construction of datasets 58
3.1.2 Computation of image features 60
3.1.3 Experiments and evaluation 61
3.1.4 Results 63
3.1.5 Discussion 65
3.1.6 Conclusion 67
3.2 ILD identification 67
3.2.1 Equal-volume lung partitioning 68
3.2.2 Computation of fractional high-density features 70
3.2.3 Experiments and evaluation 72
3.2.4 Results 73
3.2.5 Discussion 75
3.2.6 Conclusion 76
3.3 Pleural thickness measurement 76
3.3.1 Measurement using distance between pleural surface and ribs 77
3.3.2 Measurement using concave regions on the pleural surface 80
3.3.3 Experiments and evaluation 81
3.3.4 Results 82
3.3.5 Discussion 82
3.3.6 Conclusion 84

4. Image quality measurement 85
4.1 Extraction of quality measurement regions 85
4.1.1 Extraction of the EA region 86
4.1.2 Extraction of the TA region 87
4.1.3 Extraction of the AB region 88
4.1.4 Quality measurement using EA, TA, and AB 88
4.2 Evaluation methods and datasets 89
4.2.1 Phantom experiments 90
4.2.2 LDCT experiments 95
4.3 Results 97
4.3.1 Phantom experiment results 97
4.3.2 LDCT experiment results 102
4.4 Discussion 107
4.5 Conclusion 109

5. Conclusions 110
5.1 Cardiac segmentation and disease biomarker measurement 110
5.2 Lung disease biomarker measurement 112
5.3 Image quality measurement 114

References 116

LIST OF FIGURES

1.1 Estimated cancer deaths in the US in 2015....4
1.2 Simplified LCS nodule management guideline....5
1.3 Example images for LDCT....5
1.4 Example cardiac regions with different levels of noise....6
1.5 Cardiac segmentation visualization with aorta, heart region and pulmonary artery trunk....9
1.6 Examples of cardiac regions....10
1.7 Example CT scans comparing normal and early-stage ILD cases....13
1.8 Example CT scan showing an early-stage pleural thickening case....14
1.9 Automatically extracted homogeneous regions for image quality measurement....16
1.10 Dependency relationship between regions in the sequential segmentations....17
1.11 Visualization of existing segmentations in the ALM....18
1.12 Example slice in an ALM together with its intensity CT image....19
2.1 Aorta seed point detection....22
2.2 Aorta cylinder tracking model....22
2.3 Generation of cylinders with parameters in the aorta cylinder tracking stage....23
2.4 Aorta surface refinement....23
2.5 Ray averaging model and final segmentation....24
2.6 Example images of pulmonary artery....26
2.7 Two different cases before and after applying AA extension....27
2.8 Examples images showing (a)-(b) seed point detection and (c)-(d) final segmentation of PT in two cases....27
2.9 Heart region defined by constraints....28
2.10 Diagrams and examples of coronary arteries....30
2.11 Example CACs in different arteries....30
2.12 Examples of excluded high-intensity regions....31
2.13 CAC labeling model....32
2.14 Labeling of RCA CACs and LM+LAD CACs adjacent to aorta....33
2.15 Labeling of remaining left region CACs....33
2.16 Aorta and PT diameter measurement....34
2.17 Examples of AC....35
2.18 Exclusion of non-AC artifacts....36
2.19 Pericardial fat and CVF....37
2.20 CVF segmentation....38
2.21 Example reference regions for quantitative segmentation evaluation....42
2.22 Example reference regions for quantitative biomarker evaluation....42
2.23 Example cases with different visual CAC scores for the four arteries....43
2.24 Customized visualization of cardiac segmentation in coronal view (upper row) and sagittal view (lower row)....44
2.25 Axial visualization of cardiac segmentation in two slices....45
2.26 Example aorta segmentation compared to the references in 4 cases....47
2.27 Example PT segmentation compared to the references in 4 cases....47

2.28 Example CVF segmentation compared to the references in 3 cases. 48
2.29 Histogram of heart region noise on the whole dataset. 50
2.30 Bar plots showing the differences in correlation and sensitivity between the three noise groups. 51
2.31 Examples of correctly labeled CACs in a case. 51
2.32 Four example cases of AC with aorta in sagittal view 52
2.33 Examples of CAC segmentation and labeling errors. 54
3.1 Malignant and benign nodule size distribution. 60
3.2 Organization of experiment 3.1 and 3.2 63
3.3 Result ROC curves 65
3.4 Example image slices from (a) an UIP and (b) an NSIP case showing different spatial distribution of diseased lung tissues. 68
3.5 Lung partitioning visualization (3 partitions in each direction) 69
3.6 Combined partitions with 27 and 64 subsections. 70
3.7 High-density pixels in an UIP lung and a normal lung 71
3.8 The distribution of H_{right} for the same UIP and normal case shown in Figure 3.7 (27 subsections in each lung). 72
3.9 Example UIP and normal cases in the evaluation dataset 72
3.10 Histogram of H_{left} and H_{right} without partitioning for the UIP and the normal cohort. 73
3.11 Averaged h_{lefti} and h_{righti} for the UIP and the normal cohort (27 subsections in each lung) 74
3.12 Example UIP cases and normal cases. 75
3.13 Example normal cases with high h_i values. 76
3.14 Pleural thickening in two cases outlined by a radiologist. 78
3.15 Visualization of segmented and labeled ribs and lungs. 78
3.16 Ray casting model. 79
3.17 Demonstration of ray selection. 79
3.18 Examples of excluded distance measurements. 79
3.19 Visualization of final distance measurements 80
3.20 Extraction of a concave region on the pleural surface. 81
3.21 Correctly identified pleural thickening regions and normal regions incorrectly identified as pleural thickening. 83
3.22 Pleural surface regions adjacent to vertebrae in two cases with similar visual presentations. 84
4.1 Example EA regions in two cases 86
4.2 Visualization of an extracted EA region. 87
4.3 Visualization of an extracted TA region. 87
4.4 Visualization of an extracted AB region 88
4.5 Three homogeneous regions shown in a sagittal view 89
4.6 Noise and calibration profiles along the image slices 89
4.7 Average noise profiles from phantom sensitivity experiment: change of mAs (lowest and highest mAs shown). 99
4.8 Average calibration profiles from phantom sensitivity experiment: change of mAs (lowest and highest mAs shown) 100

4.9 Summary plot for phantom sensitivity experiments. 101
4.10 Summary bar plot showing mean and standard deviation for the LDCT sensitivity experiment. 103
4.11 Average noise profiles from LDCT sensitivity experiment. 104
4.12 Average calibration profiles from LDCT sensitivity experiment. 105
4.13 Noise and calibration histograms for the LCS cohort. 106

LIST OF TABLES

2.1 Aorta segmentation parameters....25
2.2 CAC segmentation parameters. 31
2.3 CAC labeling parameters. 33
2.4 AC segmentation parameters. 36
2.5 Quantitative evaluation metrics for cardiac segmentation. 39
2.6 Quantitative evaluation metrics for cardiac biomarker measurements. 40
2.7 Summary of datasets for large-scale qualitative evaluation of cardiac segmentation. 45
2.8 Quantitative evaluation results for aorta segmentation. 46
2.9 Pearson R between automated and reference scores for AS, VS, and MS. 49
2.10 Confusion matrix between automated and reference risk categories. 49
2.11 Statistics of AS scores for the whole dataset (all CACs and by artery). 49
2.12 Spearman R between visual and automated scores (all CACs and by artery). 49
2.13 Labeling sensitivity and specificity (all CACs and by artery). 50
2.14 Spearman R between visual and automated scores for the three noise groups (all CACs and by artery). 50
2.15 Labeling sensitivity and specificity for the three noise groups (all CACs and by artery). 50
2.16 Bland-Altman LOA results of diameter comparison. 52
2.17 Qualitative visual evaluation results for cardiac segmentation. 52
3.1 Nodule numbers and equivalent diameters statistics for the whole dataset and for I-ELCAP, NLST separately. 59
3.2 Nodule numbers and equivalent diameters statistics for the GA set and for each binned subset of GA. 59
3.3 Experiment 3.1 results. AUCs, IAUCs, and p-values under different training and evaluation conditions. 63
3.4 Experiment 3.2 results. Best performance for the individual bins from GA under different training and evaluation conditions. 64
3.5 Experiment 3.2 results. AUCs, IAUCs, and p-values after combining the best results from each individual bin. 64
3.6 Classification average AUC values using H_{all}. 74
3.7 Classification average AUC values using H_{max}. 74
3.8 Pearson R from pleural thickening experiments. 82
3.9 Spearman R from pleural thickening experiments. 82
3.10 Visual inspection statistics for each subsection, each lung and the two lungs. 82
4.1 Datasets summary for the longitudinal scan pairs in phantom repeatability experiment. 93
4.2 Datasets summary for phantom sensitivity experiment: change of radiation exposure. 94
4.3 Datasets summary for phantom sensitivity experiment: change of slice thickness. 94
4.4 Datasets summary for phantom sensitivity experiment: change of reconstruction

filter. 95
4.5 Datasets summary for LDCT sensitivity experiment: change of reconstruction filter. 96
4.6 Datasets summary for LDCT sensitivity experiment: change of reconstruction filter using image profiles. 97
4.7 Phantom repeatability experiment results: statistics of the differences in image quality between longitudinal scan pairs. 98
4.8 Phantom sensitivity experiment results: change in radiation exposure. 98
4.9 Phantom sensitivity experiment results: change in slice thickness. 98
4.10 Phantom sensitivity experiment results: change in reconstruction filter. 98
4.11 LDCT repeatability experiment results: statistics of the differences in image quality between longitudinal scan pairs. 102
4.12 LDCT sensitivity experiment results: change in reconstruction filter. 102
4.13 LDCT experiment results: statistics of quality on a LCS cohort. 106

CHAPTER 1

INTRODUCTION

Computer-aided diagnosis (CAD) is the application of computer methods to aid physicians in the analysis of medical images. In this dissertation, novel fully-automated CAD algorithms have been developed to perform thoracic organ segmentation and quantitative biomarker measurement from chest CT images with a focus on the low-dose CT images associated with lung cancer screening. The CAD algorithms in this dissertation can be grouped into three categories:

1) Algorithms to segment cardiac organs and measure cardiac disease image biomarkers. The cardiac organs include the thoracic aorta, the pulmonary artery trunk, and the general heart region. The cardiac biomarkers include the coronary artery calcification (CAC) content, the aorta and pulmonary trunk (PT) diameter profile, the cardiac visceral fat content, and the aortic calcification (AC) content. Segmentation of cardiac organs take advantage of other adjacent pre-segmented organs such as the lungs and the fat tissues. The segmented cardiac organs are used as masks for the segmentation and measurement of cardiac biomarkers.

2) Algorithms to characterize three different types of lung region image biomarkers: the pulmonary nodule, early-stage interstitial lung disease (ILD), and early-stage pleural thickening. For the pulmonary nodule, an algorithm is developed to characterize nodule malignancy using image features. Furthermore, different datasets and experiments have been constructed to evaluate the impact of nodule size on characterization outcome. For ILD, an algorithm has been developed to partition the lung region into equal-volume subsections and a fractional high-density feature is computed for each subsection to indicate the spatial distribution of ILD related lung tissues. For pleural thickening, two different algorithms have been developed to detect two

different types of pleural thickening, by measuring the distance between the pleural surface and the ribs and by detecting concave regions on the pleural surface.

3) An algorithm to assess image quality (noise and calibration level) from 3D homogeneous regions in CT images. The 3D homogeneous regions are computed using pre-segmented organs such as the trachea and the descending aorta (DA). An external air region is also computed. Noise is represented as the pixel standard deviation in a homogeneous region and calibration is represented as the mean intensity level. Both the global image quality and a per image slice quality profile are computed for all three regions.

In the following sections, section 1.1 describes the primary target dataset used in this dissertation: the low-dose chest CT scans associated with lung cancer screening; section 1.2 gives literature reviews and discusses technical challenges associated with each main topic of the dissertation: cardiac analysis, lung region analysis and image quality assessment; section 1.3 describes the foundation of all the CAD algorithms in this dissertation: the pre-segmentation of other chest organs; section 1.4 gives an overview of the subsequent chapters in the dissertation.

1.1 Low-dose chest CT from lung cancer screening

The annual screening for lung cancer with low-dose computed tomography (LDCT) has been approved for reimbursement for high-risk population [1] in 2015. Based on the Centers for Medicare & Medicaid Services summary, the eligibility criteria of lung cancer screening (LCS) participants include age 55 to 77 years (55 to 80 years recommended by the Unites States Preventive Services Task Force (USPSTF) [2]), asymptomatic, tobacco smoking history of at least 30 pack-years and current smoker or one who has quit smoking within the last 15 years. Using the USPSTF eligibility criteria, Jemal et al. [3] estimated that in 2015, 6.8 million smokers were eligible for LDCT screening. However, they have also found that currently only a small

percentage of eligible smokers have reported LDCT screening (3.9% in 2015). In the future, there will potentially be many more LDCT scans taken annually through LCS.

Although the primary purpose of LCS is to detect early-stage pulmonary nodules, the availability of such a large dataset of whole chest CT scans on an annual basis also provides the opportunity for screening for other chest region diseases. Due to the age range and smoking history, these LCS participants are typically at high risk of heart diseases [4]. Therefore, it is beneficial to provide companion diagnosis using these LDCT scans such as evaluating CAC content. The companion diagnosis offers additional disease risk evaluation beyond the detection of pulmonary nodules without exposing the participants to additional radiation.

This dissertation builds upon the research studies and LDCT datasets developed and gathered over the past 20 years for LCS. These datasets and studies have enabled the development and large-scale evaluation of robust CAD algorithms presented in this dissertation. In the following subsections, section 1.1.1 will give more details on the benefits of annual LCS and section 1.1.2 will discuss the challenges associated with developing CAD algorithms for LDCT.

1.1.1 Lung cancer and lung cancer screening

Lung cancer causes more deaths than any other type of cancer in the US [5]: it is estimated that in 2015, 158,040 deaths were caused by lung and bronchus cancer. In fact, lung cancer caused about as many deaths as the next four deadliest cancers combined together: colon, pancreas, breast and prostate cancers as shown in Figure 1.1.

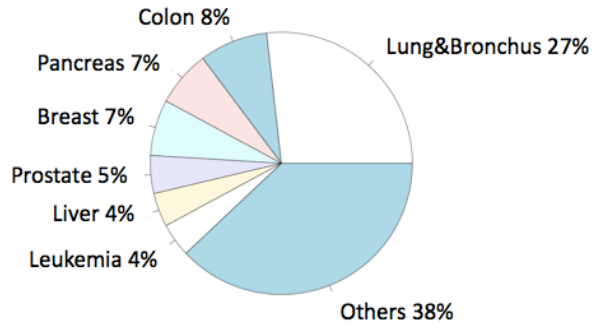


Figure 1.1: Estimated cancer deaths in the US in 2015.

It has been established that screening for lung cancer greatly improves lung cancer survival rate [6]. The International Early Lung Cancer Action Program (I-ELCAP) investigators have shown that among the screening participants who received a diagnosis of lung cancer, 85% had clinical stage I lung cancer, the stage where cancers are no larger than 3cm and are confined to the lungs [7]. The estimated 10-year survival rate for this subgroup was 88% regardless of treatment. Of this stage I lung cancer group, the participants who underwent surgical resection within 1 month after diagnosis had an estimated 10-year survival rate of 92%. Compared to the much lower survival rate for the general population without screening (17.7% 5-year) [8], lung cancer screening greatly increases survival rate. The majority of the lung cancer cases from screening are stage I lung cancer, the only stage at which cure by surgery is highly likely.

Screening for lung cancer has been approved for insurance reimbursement for eligible high-risk population in 2015 by the Centers for Medicare & Medicaid Services [1] (eligible criteria include 55-77 years, asymptomatic and tobacco smoking history of at least 30 pack-years). The screening process involves a baseline LDCT scan. If no nodule is detected, the participant takes annual repeat LDCT scans. If nodules are detected, depending on the nodule size and solidity, different follow-up procedures will be taken [9]. Although LCS nodule management guidelines vary among institutes and associations, the general guideline can be simplified as shown in Figure 1.2.

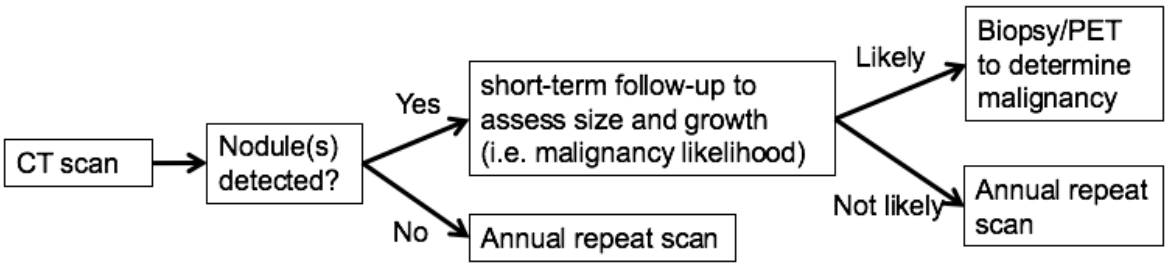
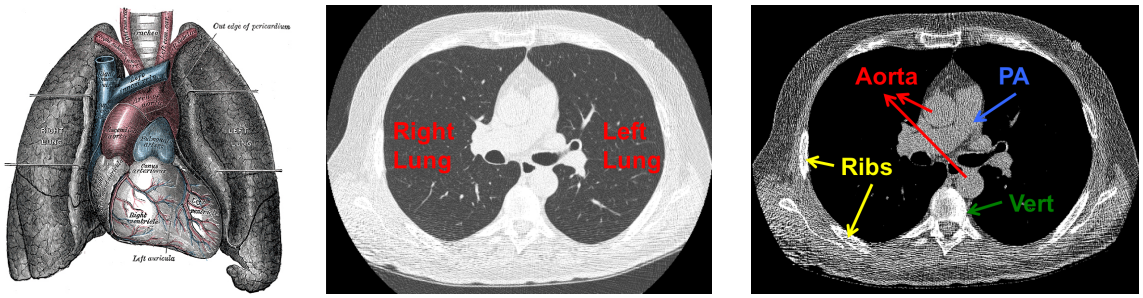


Figure 1.2: Simplified LCS nodule management guideline.

1.1.2 Characteristics of low-dose CT

The LDCT involved in LCS is typically a thoracic CT scan that covers the whole lung region [10]. It usually starts a few millimeters superior to the lungs and ends 1-2 centimeters inferior to the lungs. The entire heart region and pulmonary arteries are also covered, providing the opportunity to assess additional disease risk factors such as CAC content. Figure 1.3 (a) shows an illustration of the anatomical structures in the chest region including lungs, heart and major vessels [11]. Figure 1.3 (b) and (c) shows an axial image slice from a LDCT scan using different image window settings. Different anatomical structures are also indicated in Figure 1.3 (b) and (c) including the lungs, the aorta, and the main pulmonary artery.

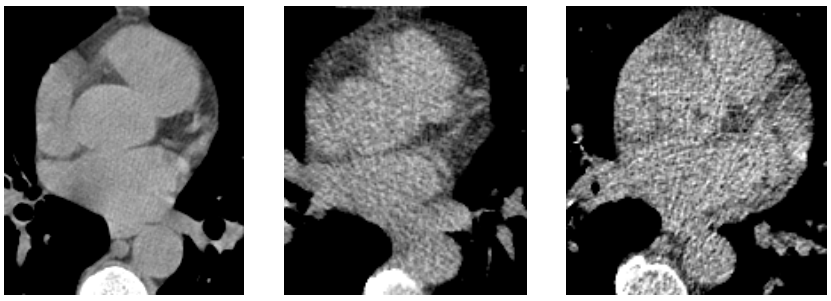


(a) (b) (c)

Figure 1.3: Example images for LDCT. (a) illustration of anatomical structures included in LDCT; (b) LDCT image slice using lung window; (c) LDCT image slice using abdominal window. The pulmonary artery is labeled as PA and the vertebra is labeled as Vert.

In a similar manner to traditional clinical CT, LDCT is calibrated, and homogeneous regions fall into pre-determined Hounsfield Units (HU) ranges [12]. However, the radiation dose level is much lower for LDCT compared to clinical CT scans. Ono et al. (2013) [13] reported

that the effective dose of LDCT was between 1.3 and 3.4 millisievert (mSv) while that of the standard-dose CT was between 8.5 and 14.0 mSv. The USPSTF has reported that the radiation exposure associated with LDCT ranges from 0.61 to 1.5 mSv per scan [2]. The lower radiation dose results in higher image noise. Vardhanabhuti et al. (2013) [14] compared the image noise level of standard and low-dose chest CT using different reconstruction schemes and have found that when the mean effective dose was 3.7 mSv, the filtered back projection (FBP) reconstructed scans had a mean image noise of 26.9 HU and when the dose was decreased to 0.9 mSv, the FBP scans' image noise increased to 51.8 HU. Figure 1.4 shows example images in datasets used in this dissertation. The cardiac region noise is 22 HU in the standard-dose image shown in (a) and 35-48 HU for the low-dose images shown in (b) and (c). Therefore, most established computer algorithms designed for standard-dose CT analysis do not translate well to low-dose CT images. Low-dose CT image analysis requires novel noise-robust approaches and methods.



(a) (b) (c)

Figure 1.4: Example cardiac regions with different levels of noise. (a) standard-dose CT (aorta noise=22HU); (b) low-dose CT (aorta noise=35HU); (c) low-dose CT (aorta noise=48HU).

1.2 Computer aided diagnosis in low-dose CT: literature review and challenges

CAD is playing an increasingly important role in medical research and clinical practice. Most of the quantitative CAD systems today are semi-automated, requiring user interaction such as manual annotation of a seed region or manual correction of computer outcomes. Fully-automated CAD does not produce intra-observer variations and, if well-designed, can be more sensitive and efficient in the detection of subtle structures easily missed by human such as

pulmonary nodules [15]. Zhao et el. (2012) [15] compared the performance of double reading versus CAD in the detection of pulmonary nodules in LDCT and have found that the sensitivity of nodule detection was 78.1% for double reading and 96.7% for CAD. Beyer et al. (2007) [16] compared the reading time of pulmonary nodule detection from CT scans using CAD as a concurrent reader. They have found that reading time was significantly shorter for using CAD as a concurrent reader (mean reading time 274s per case) compared to reading without CAD (mean reading time 294s per case) without significant difference in sensitivity for nodules size 3mm and above.

Fully-automated CAD is the focus of this dissertation. Compared to semi-automated approaches, full-automation is expected to have a more efficient workflow without the need of manual adjustments or corrections, therefore providing the foundation for higher quality healthcare. All segmentation and biomarker measurement algorithms presented in this dissertation are fully-automated and only require the CT scan and its Anatomy Label Map as inputs. The Anatomy Label Map will be introduced and defined in section 1.3.

The three major topics of this dissertation, which will be reviewed in the following subsections, are:

1) fully-automated computer algorithms for the segmentation of cardiac organs and measurement of cardiac disease related image biomarkers from LDCT;

2) fully-automated computer algorithms for the measurement of lung disease related image biomarkers from LDCT;

3) a fully-automated computer algorithm to assess image quality (noise and calibration) from homogeneous regions in CT scans.

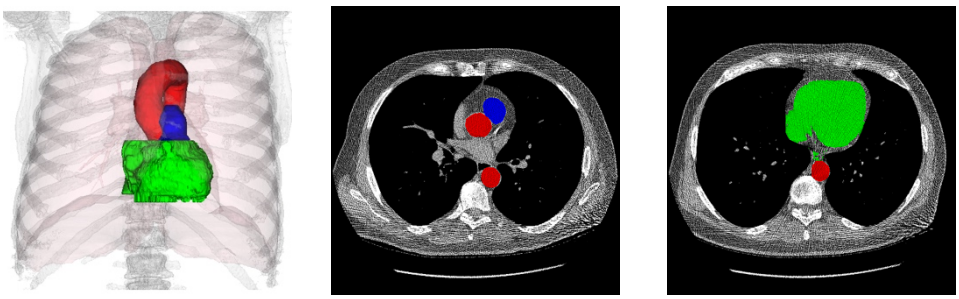
In this section, relevant state-of-the-art CAD algorithms are reviewed with a special focus

on those related to LDCT. Section 1.2.1 reviews cardiac region CAD algorithms; section 1.2.2 reviews lung region CAD algorithms; section 1.2.3 reviews image quality assessment algorithms.

1.2.1 CAD in the cardiac region

CAD analysis in the cardiac region typically follows two steps: first, the relevant cardiac organs are segmented; then, based on the segmented regions, image biomarkers are measured such as the heart volume or the amount of calcium each coronary artery contains. The automated cardiac CAD algorithms in this dissertation also follow these two general steps: first cardiac organs are segmented using spatial modelling and constraints provided by other adjacent pre-segmented organs such as the lungs, bones and fatty tissues; then cardiac image biomarkers are measured based on the segmented organs. The CAD algorithms are designed for LDCT scans although they are also applicable to standard-dose non-contrast chest CT scans.

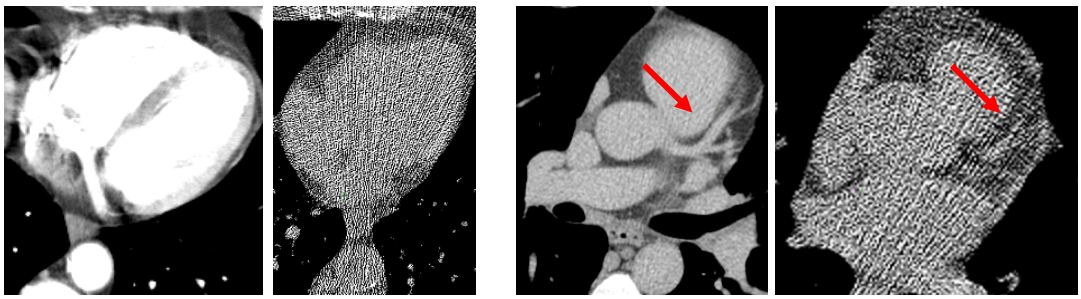
Cardiac segmentation starts from the well-defined aorta region: cylinder tracking is used to determine the approximate aorta location and ray casting based surface refinement is used to determine the final aorta surface. The segmented aorta then provides constraints to the PT and heart region. The PT is segmented using a method similar to that used for the aorta. The heart region is defined based on constraints from adjacent organs. Then the segmented heart region is used as a mask to extract CAC. Based on the segmented organs, biomarkers are automatically measured such as the CAC content and aorta diameter profile. A complete cardiac segmentation visualization is shown in Figure 1.5.



(a) (b) (c)

Figure 1.5: Cardiac segmentation visualization with aorta (red), heart region (green) and pulmonary artery trunk (blue). (a) coronal view; (b) and (c) show axial slices.

Traditionally cardiac segmentation and measurement algorithms are often designed for CT angiography (CTA) scans, gated cardiac CT scans, or contrast-enhanced CT scans because in these scans the cardiac region is highly visible and the boundaries can be detected with little ambiguity. For instance, Ecabert et al. (2008) [17] performed fully-automated heart chamber segmentation in contrast-enhanced CT using a model-based approach. Kelm et al. (2011) [18] performed the fully-automated detection of coronary stenosis in CTA scans by estimating coronary vessel location and size and performing classification. Wolterink et al. (2015) [19] automatically segmented and labeled CAC in ECG-triggered cardiac CT scans using feature classification. However, in the context of LDCT, cardiac regions such as the heart and coronary arteries do not have clearly defined boundaries and cannot be robustly segmented. Figure 1.6 (a) shows a contrast-enhanced CT image heart region where the individual heart chambers can be clearly seen and Figure 1.6 (b) shows a heart region in a LDCT scan where only the general heart fat transition boundary is visible. Furthermore, LDCT from LCS are not ECG-gated, which further increases the difficulty in segmenting small structures such as the coronary arteries or extracting meaningful image features. Figure 1.6 (c) shows a clearly-defined coronary artery in a gated standard-dose CT scan while the coronary artery in the same location in a LDCT has a more blurred boundary as shown in Figure 1.6 (d).



(a) (b) (c) (d)

Figure 1.6: Examples of cardiac regions. (a) contrast-enhanced CT heart region; (b) LDCT heart region; (c) standard-dose gated CT scan coronary artery (red arrow); (d) LDCT coronary artery (red arrow).

Although the high noise level and heart motion artifacts associated with LDCT impact the precision of image biomarker measurement, it has already been shown that cardiac biomarkers detected from LDCT such as CAC content are clinically useful. Shemesh et al. (2006) [20] have shown that CAC score derived from ungated LDCT can contribute to risk stratification and management of coronary artery disease. Shemesh et al. (2010) [21] have also shown that visual assessment of CAC on LDCT scans provides clinically relevant quantitative information as to cardiovascular death. LDCT related fully-automated cardiac algorithms have been developed by Isgum et al. (2009, 2012) [22-23]. In their work, the aorta and heart region were segmented using a multi atlas-based registration approach and the CAC was detected using a probability map and image feature-based classification.

1.2.2 CAD in the lung region

There are a variety of CAD related studies concerning lung health diagnostics including the detection and characterization of pulmonary nodules and lung parenchymal diseases. In this section, only studies related to this dissertation are reviewed. They are:

1) the characterization of small pulmonary nodules;

2) the detection and characterization of early-stage ILD;

3) the detection and characterization of early-stage pleural thickening.

1) Characterization of small pulmonary nodules

In this dissertation, an algorithm has been developed to characterize small pulmonary nodule malignancy status and to evaluate the impact of nodule size distribution on characterization outcome. A large dataset with a total of 736 LCS nodules is used for training and evaluation in this study (size range 3 to 30 mm). A subset of size-balanced nodules is constructed which contains 163 malignant and 163 benign nodules of the same size distribution (size-range 5 to 14 mm). This subset is used to evaluate the impact of nodule size distribution as well as to explore the optimal training strategies when evaluating on a size-balanced nodule dataset.

Most previous studies on nodule characterization by other research groups involve clinically detected nodules such as Shah et al. (2005) [24] (35 nodules, diameter ranged from 6 to 54 mm with a mean of 25 mm) and Way et al. (2006) [25] (96 nodules, longest diameter ranged from 3.9 to 59.8 mm with a mean of 17.3 mm). These nodules are not very relevant in the context of LCS because the maximum nodule diameter should not exceed 30 mm in LCS [26]. A small number of studies used nodules from LCS (Suzuki et al. 2005, Aoyama et al. 2003, Armato et al. 2003) [27-29]. These three studies have all used the same dataset with more than 400 benign nodules and less than 80 malignant nodules. Nodule size distribution was not taken into consideration in these three studies and the benign nodules were smaller in size than the malignant ones.

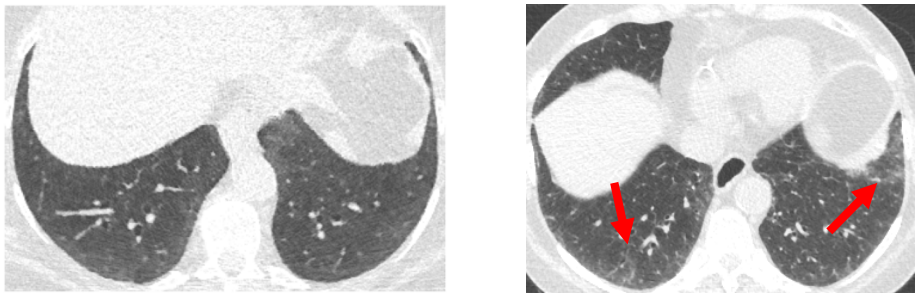
Pulmonary nodule size is an important indicator for nodule malignancy. Our group's previous work has indicated that ignoring nodule size could lead to overly-optimistic characterization result (Artit et al. 2007) [30]. In medical studies, Wisnivesky et al. (2004) [31] have reported that for stage I non-small cell lung cancers, the most curable cancers are below 15

mm in diameter and the larger they were, the less curable they became. A similar observation was made by McWilliams et al. (2013) [32]. They have found that the median size of the benign nodules was 3 to 3.4 mm while that of the malignant nodules was 12.8 to 13 mm in a LCS study. Because larger nodule is more likely to be malignant, in this dissertation a subset of malignant and benign nodules of the same size range (size-balanced nodules) has been constructed to evaluate their image feature difference independent of size. This is the first time a nodule characterization study has been performed on a large size-balanced dataset consisting of nodules less than 15 mm, the most relevant size range for LCS.

2) Characterization of early-stage ILD

ILD is a group of lung parenchymal diseases characterized by progressive scarring of the lung tissue [33]. Some types of ILDs have known causes such as the exposure to allergens while others have unknown causes (idiopathic). ILD is generally irreversible and therefore it is important to detect ILD in its early stages.

In this dissertation, a fully-automated algorithm has been designed to capture early-stage ILD from LDCT scans. In early-stage ILD, the abnormal tissue patterns are relatively localized and do not exhibit very distinguishable texture patterns. Figure 1.7 shows a normal CT scan together with an early-stage ILD scan. Rather than relying on complex texture features, the algorithm partitions the lung region into equal volume subsections and extracts a classic high-density feature from each subsection to characterize the spatial distribution of parenchyma fibrosis. This feature vector is computed for both ILD cases and normal cases and is used for the classification of ILD versus normal CT scans.



(a) (b)

Figure 1.7: Example CT scans comparing normal and early-stage ILD cases. (a) a normal case; (b) an early-stage ILD case (red arrows point to the lung fibrosis regions).

In other previous studies, image patches were manually cropped from lung regions and image feature based classification was performed on these image patches. Depeursinge et al. (2012) [34] classified lung parenchymal patches into normal, emphysema, ground glass, fibrosis, and micronodules using wavelet related features. Xu et al. (2006) [35] calculated volumetric features including statistical, histogram, and fractal features from manually drawn volumes of interest in the lung parenchyma and classified the volumes into emphysema, ground-glass, honeycombing, normal nonsmokers, and normal smokers using these volumetric features. Korfiatis et al. (2010) [36] classified manually defined volumes of interest in the lung parenchyma into normal, ground glass, and reticular patterns using 3D co-occurrence features. The algorithm developed in this dissertation is distinctive from other research studies (Depeursinge et al. 2012, Xu et al. 2006, Korfiatis et al. 2010) [34-356] in two major aspects. First, the characterization target is the entire CT scan not lung parenchymal image patches. Second, it does not rely on complicated feature designs since the fibrosis is still early-stage and does not form unique texture patterns. Instead, it focuses on the detection of the unique spatial distribution of lung fibrosis through spatially partitioning the lung regions.

3) Characterization of early-stage pleural thickening

Pleural thickening is the fibrosis of the visceral pleura. It is a lung disease associated with asbestos exposure (Gevenois et al. 1998) [37]. Conventionally, visual assessment is applied to

advanced pleural thickening such as those observed in patients with malignant pleural mesothelioma. In this dissertation, the target disease is early-stage pleural thickening as shown in Figure 1.8. It is much more subtle than advanced stage pleural thickening and should be characterized using different, more sensitive, methods.

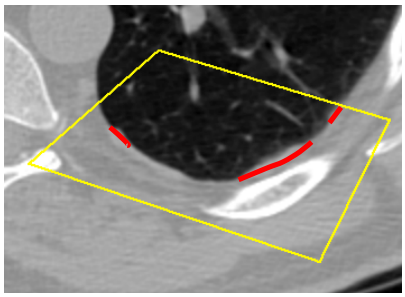


Figure 1.8: Example CT scan showing an early-stage pleural thickening case. Red lines outline the abnormal pleural surface region.

In this dissertation, two different fully-automated pleural thickening characterization methods are presented to address the different manifestations of pleural thickening. The first method captures the subtler pleural thickening by computing the distance between pleural surface and ribs. This distance is used as a surrogate for the actual pleural thickness and it is hypothesized that for an early-stage pleural thickening case, this distance will be larger than that of a normal case. This method is designed to measure the pleural thickening regions not easily detectable visually. The second method captures the more conventional and visually detectable pleural thickening regions by detecting the concavities on the costal pleural surface. The volume of the concavities represents the pleural thickening severity of a given case.

In other previous studies, pleural thickening is usually detected through two steps. First the abnormal regions on the lung surface are extracted often based on shapes. Then image feature classification is used to classify the extracted regions into pleural thickening and others. For instance, Sensakovic et al. (2011) [38] used a semi-automated method to segment the lung parenchyma pleural space. Then a classifier was used to identify malignant pleural mesothelioma

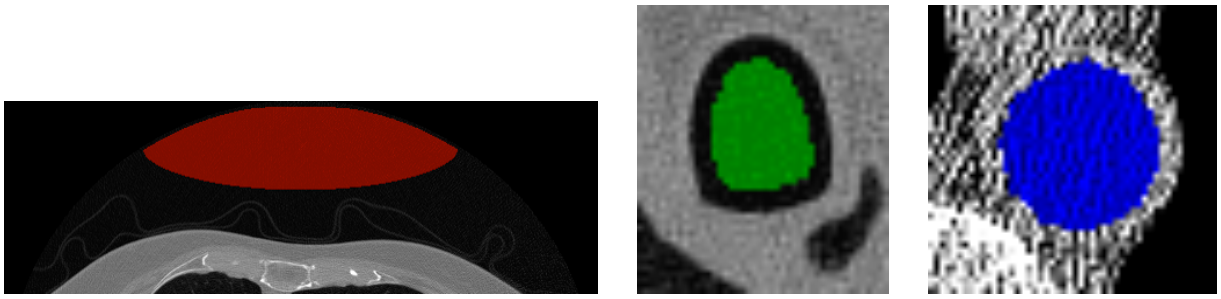
in the pleural space. Rudrapatna et al. (2005) [39] extracted pleural thickening candidates and computed pixel-based features to classify candidate pixels into pleural plaque or not plaque. Chaisaowong et al. (2014) [40] delineated pleural contours and obtained the pleural thickening regions using 3D tissue features. However, the automated studies were evaluated on relatively small datasets (140 image slices for Rudrapatna et al. [39] and 27 image slices for Chaisaowong et al. [40]) and their robustness on larger datasets is unknown.

1.2.3 CT image quality assessment

The automated assessment of CT image quality is a necessary first step to ensure CT images are acceptable to be used in visual and automated analysis. CT image quality such as noise and calibration level is especially important to automated CAD algorithms since these algorithms are developed based on the assumptions that the images are well calibrated and have an acceptable noise level. For example, a CAD algorithm developed for images with a moderate noise level may not translate well to images with very high noise levels.

In this dissertation, a fully-automated method, which has been developed to measure CT image quality including noise and calibration values from 3D homogeneous regions in chest CT, is presented. This algorithm evaluates three 3D homogeneous regions: an external air region (EA), a trachea lumen region (TA) and a DA lumen blood region (AB) (see Figure 1.9). All measurement regions are automatically computed from existing segmentations and are in 3D space. The EA region spans through the extent of the whole lungs; the TA region starts from the top of the lungs and extends to the airway carina; the AB region starts at the airway carina and extends until the end of the segmented DA. The global noise and calibration values are measured in each homogeneous region. In addition, a per-slice quality profile is also computed for each region, i.e. the calibration and noise are measured for each axial slice to form a per-slice quality

profile spanning the whole lung region. This quality measurement method is validated on large datasets with various imaging parameters to demonstrate its repeatability and sensitivity.



(a) (b) (c)

Figure 1.9: Automatically extracted homogeneous regions for image quality measurement. (a) external air region (red); (b) airway lumen (green); (c) aorta lumen (blue).

Previous studies measuring CT image quality have all used manually defined 2D circular image regions. Singh et al. (2011) [41] used manually drawn circular Region of Interests (ROIs) inside DA to measure image noise and calibration values. Ohno et al. (2012) [42] placed manually drawn circular ROIs inside trachea lumen, lung parenchyma, aorta, heart and other homogeneous regions for noise measurement. Yamada et al. (2012) [43] drew circular ROIs in the trachea, the pectoralis major muscle and the aorta to compute signal to noise ratio. In contrast, the method in this dissertation does not require manual markings and is able to characterize both the global image quality and the quality spatial variations. It should be noted that the quality assessment method presented in this dissertation measures image quality from the reconstructed CT images retrospectively and is therefore fundamentally different from the dose modulation method used clinically to minimize radiation dose while maintaining acceptable image quality [2, 44].

1.3 Segmentation using the Anatomy Label Map

The previous sections have given an overview of the main algorithms that are presented in the subsequent chapters. All these algorithms share a common characteristic: they utilize the pre-segmentation of other organs. For instance, the cardiac segmentation relies on the segmented

fat tissues, lung regions and bone structures. The ILD characterization uses the pre-segmented lung regions and the pleural thickening detection uses both the segmented lungs and labeled ribs. These pre-segmentations are performed using algorithms previously developed by our research group. These pre-segmented organs and regions include: the airway tree with labeled trachea and main bronchi (Lee et al. 2009) [45], the left and right lungs (Kostis et al. 2003) [46], skin surface and fat tissue (Padgett et al. 2014) [47], labeled individual ribs (Lee et al. 2010) [48] and other bone structures such as the sternum (Liu et al. 2015) [49] and the vertebra (Liu et al. 2016) [50]. In this subsection, a general overview of the segmentation framework and the pre-segmented regions is given.

In our segmentation system, a top-down strategy is used where segmentation starts from the most robust and simplest regions. Then subsequent segmentations take advantage of the previously segmented regions (Reeves et al. 2017) [51]. The dependency relationship used in the sequential segmentations is shown in Figure 1.10.

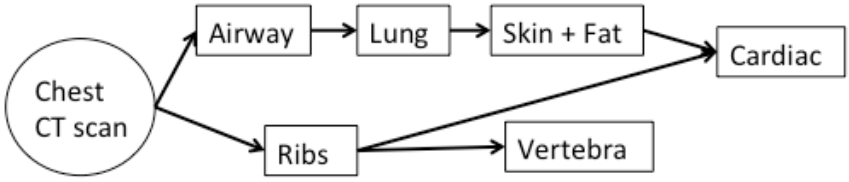


Figure 1.10: Dependency relationship between regions in the sequential segmentations.

The airway tree is segmented using a cylinder tracking and 3D region growing based method (Lee et al. 2009) [45]. In the segmented airway tree, the trachea and main bronchi are labeled. The trachea branching point is determined to be the carina. The colored airway tree segmentation is shown in Figure 1.11 (a).

The left and right lungs are segmented using image filtering, intensity thresholding and morphological operations (Kostis et al. 2003) [46]. The left and right lungs are partitioned with a minimum distance path-cutting algorithm. The colored two lungs are shown in Figure 1.11 (b).

The skin surface is segmented by differentiating body from the outside air region. Fat is segmented using a local noise-aware algorithm (Padgett et al. 2014) [47]. The segmented skin and fat regions are shown in Figure 1.11 (c).

The individual ribs are segmented by employing an algorithm based on region growing and cylinder tracking (Lee et al. 2010) [48]. The labeled individual ribs are shown in different colors in Figure 1.11 (d).

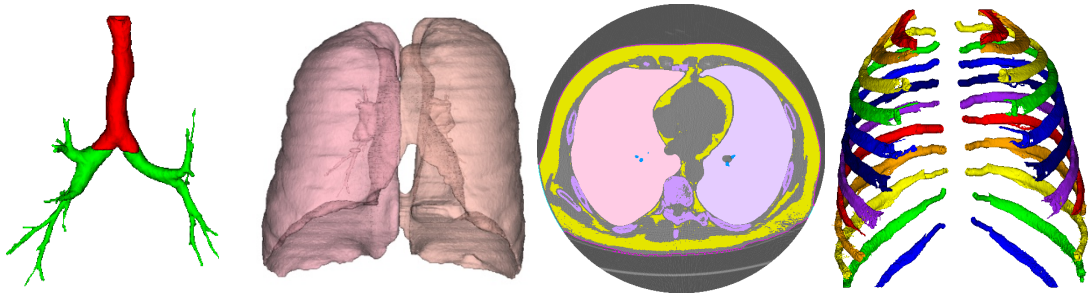
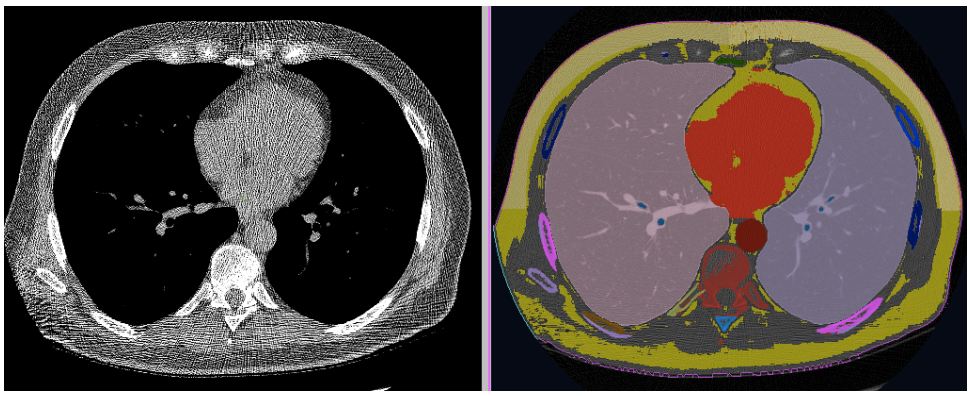


Figure 1.11: Visualization of existing segmentations in the ALM. (a) segmented airway tree with trachea (red) and bronchi (green); (b) segmented left and right lungs in different shades of pink; (c) segmented skin surface (magenta) and fat (yellow); (d) segmented and labeled individual ribs in different colors.

Once a region is segmented, the segmentation result will be recorded into an Anatomy Label Map (ALM). The ALM is of the same size as the CT image and pixels corresponding to the segmented regions are labeled based on the region they are associated with. For example, in Figure 1.12, one axial slice of the ALM is shown together with the intensity image and each different label in the ALM is shown in a different color: fat pixels in yellow, left lung and right lung in two different shades of pink, ribs in multiple colors, skin surface in magenta, aorta in dark red, heart region in bright red.



(a) (b)

Figure 1.12: Example slice in an ALM together with its intensity CT image. (a) intensity CT; (b) ALM where each label is shown in a unique color.

Based on the dependency relationship in Figure 1.10, a later segmentation is not allowed to overwrite an earlier segmentation. For instance, lung segmentation is not allowed to include pixels already labeled as airway and cardiac segmentation is not allowed to include pixels already labeled as bone. The use of ALM and the sequential segmentation mechanism ensures that the more complex segmentations are built upon robust foundations.

1.4 Overview

The following sections are organized as follows: Chapter 2 presents the cardiac organ segmentation and biomarker measurement algorithms, which includes the segmentation of aorta, PT, the general heart region and the segmentation of CAC, AC and cardiac visceral fat. Measurement of calcification content, aorta and PT diameters is also described. Chapter 3 presents the lung region analysis which includes the characterization of pulmonary nodule, the detection and characterization of early-stage ILD and early-stage pleural thickening. Chapter 4 presents an automated algorithm to assess CT image quality (noise and calibration) from homogeneous regions. Chapter 5 presents the conclusion.

CHAPTER 2

CARDIAC SEGMENTATION AND DISEASE BIOMARKER MEASUREMENT

Heart disease is the leading cause of deaths for both men and women. In the US, about 610,000 people die from heart diseases every year [4]. Among all heart diseases, coronary heart disease is the most common type, killing about 365,000 people in 2014 [4]. Key risk factors for heart diseases include high blood pressure, high cholesterol, and smoking. Therefore, the subjects enrolled in lung cancer screening (aged 55-80 with a history of smoking [2]) are typically also at high risk of heart diseases. Evaluating heart disease risk factors on lung cancer screening CT scans offers additional benefit without exposing the subject to additional radiation.

In order to evaluate an individual's heart disease risk factor, the cardiac organs are segmented and disease related biomarkers are measured. In the cardiac region, aorta is very well-defined: it originates from the heart and extends through the thoracic region; it also has a cylindrical shape and a relatively clear boundary. Therefore, cardiac segmentation starts from the aorta, based on which other cardiac regions are obtained.

The most important cardiac biomarker is the amount of CAC, an indicator of coronary heart diseases. This biomarker is obtained and measured using the spatial constraints provided by other segmented regions. Other biomarkers such as the aorta diameter profile and cardiac fat content are also measured. The following subsections are arranged as follows:

1) Section 2.1 describes the cardiac organ segmentation algorithms: aorta in section 2.1.1, pulmonary artery trunk in section 2.1.2, and heart region in section 2.1.3;

2) Section 2.2 describes the cardiac biomarker segmentation and measurement algorithms: CAC in section 2.2.1, aorta and PT diameter in section 2.2.2, AC in section 2.2.3, and cardiac visceral fat in section 2.2.4;

3) Section 2.3 presents the quantitative and qualitative evaluation methods: section 2.3.1 describes the manual marking procedures and quantitative evaluation experiments; section 2.3.2 describes the qualitative evaluation experiments;

4) Section 2.4 presents the results for both quantitative (section 2.4.1) and qualitative (section 2.4.2) evaluations;

5) Section 2.5 and 2.6 are the discussions and conclusions.

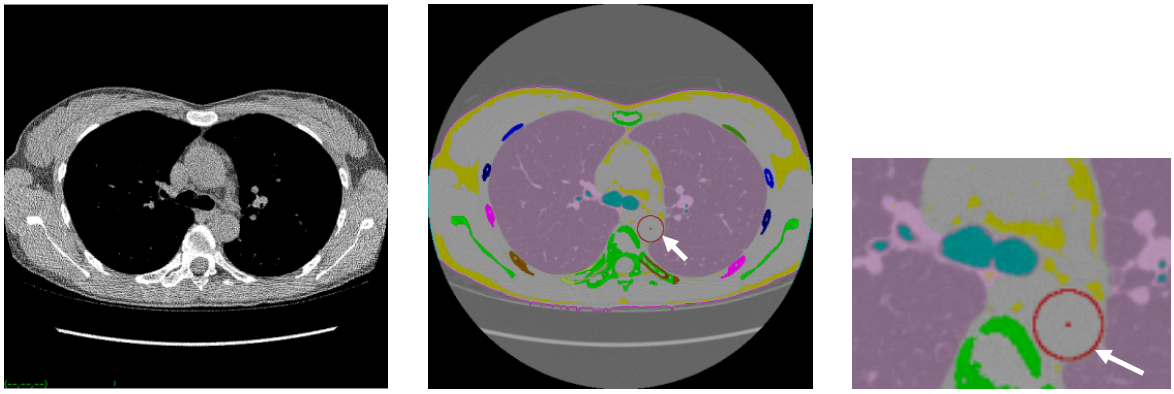
2.1 Segmentation of the aorta, the pulmonary artery trunk and the heart region

In this algorithm, first the aorta is segmented using a three-stage approach. Then the PT region is located using constraints from the ascending aorta (AA) segmentation. Finally, the general heart region is segmented based on constraints from the segmented aorta, the PT, as well as other anatomical structures. All segmentations utilize constraints from the ALM, which contains pre-segmented organs including the major airways, the lungs, ribs, and fatty tissues. The segmentation algorithm details for each of these three regions are described in the following subsections.

2.1.1 Segmentation of the aorta

A novel three-stage algorithm (Xie et al. 2014) [52] is used to segment the aorta. In the first stage a seed point central to the aortic lumen is identified by circle fitting in the axial image slice at the level of the carina. Then a cylinder tracking algorithm is used to track the approximate aorta centerline upward and downward starting from the seed point. Finally, a surface refinement method is used to obtain the final aorta surface. The details of each of these stages are given in Algorithm 2.1 to Algorithm 2.3. Seed point detection is illustrated in Figure 2.1. Figure 2.1 (a) shows the axial image slice at the level of the carina. Figure 2.1 (b) and (c) show the ALM for that image slice as well as the detected seed point (center of circle). Figure

2.2-2.3 show the model and details of the cylinder tracking stage and Figures 2.4-2.5 show the surface refinement stage.

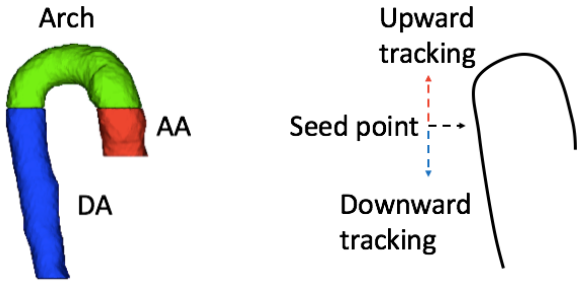


(a) (b) (c)

Figure 2.1: Aorta seed point detection. (a) intensity CT image at the level of the carina; (b) ALM with pre-segmented organs in different colors; (c) an enlarged view of (b). Each color corresponds to a label: lungs (pink), fat (yellow), skin (magenta), airway carina (sea green), ribs (multi-color), vertebra and other bones (light green). The red circle (indicated by the arrow) is the largest circle fit into the unlabeled region.

Algorithm 2.1 Aorta seed point location

1) Select the axial label map at the level of the carina M_c (see Figure 2.1 (b));
2) Identify a ROI I_s in M_c adjacent to the segmented vertebra I_{vert} and the segmented left lung I_{llung}. Obtain the unlabeled region I_e, where
$$I_e = I_s \cap \bar{I}_{vert} \cap \bar{I}_{llung} \quad (2.1)$$
3) Denote a circle C as (s_x, s_y, r_c) where (s_x, s_y) is the center of C and r_c is its radius. Find the largest C:
$$\underset{(s_x,s_y)}{\operatorname{argmax}}\, r_c = \{(s_x, s_y) | C = (s_x, s_y, r_c) \in I_e\} \quad (2.2)$$
The corresponding (s_x, s_y) is the seed point (see Figure 2.1 (b)-(c));



(a) (b)

Figure 2.2: Aorta cylinder tracking model. (a) an aorta divided into ascending (AA in red), arch (Arch in green) and descending (DA in blue) sections; (b) upward and downward tracking model.

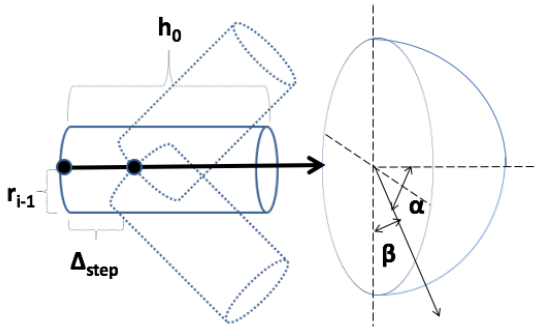
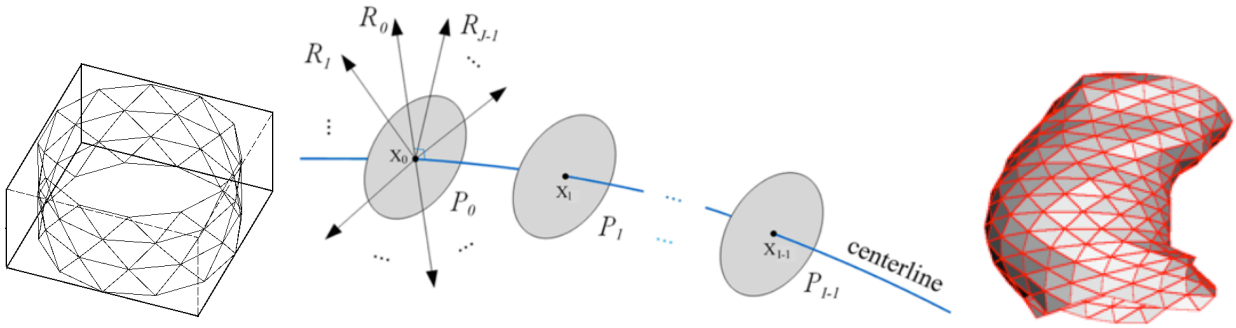


Figure 2.3: Generation of cylinders with parameters in the aorta cylinder tracking stage.

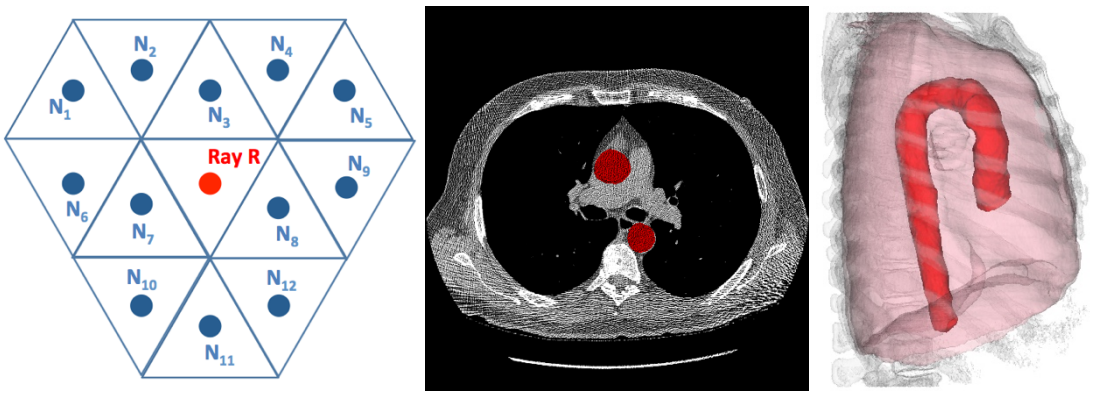
Algorithm 2.2 Cylinder tracking algorithm

1) From the seed point (s_x, s_y), successively fit cylinders into the unlabeled region upwards and downwards (see Figure 2.2). Each cylinder Cyl is denoted by its length h, angle orientation (α, β) and radii r. The set of cylinders in the first iteration is Cyl_0, where
$$Cyl_0 = \{(h_0, r_0, \alpha_0, \beta_0) | \alpha_0 \in \alpha, \beta_0 \in \beta\} \quad (2.3)$$
2) Denote the number of unlabeled pixels in Cyl_0 as M_1 and the number of labeled pixels in Cyl_0 as M_2. Compute a matching score function:
$$S_0 = M_1 - \omega * M_2 \quad (2.4)$$
where ω is the weight coefficient. Select the cylinder Cyl_{0M} with the maximum score as the best matching cylinder:
$$\underset{Cyl_{0M}}{\operatorname{argmax}} S_0 = \{Cyl_{0M} = (h_{0M}, r_{0M}, \alpha_{0M}, \beta_{0M}) | h_{0M} = h_0, r_{0M} = r_0, \alpha_{0M} \in \alpha, \beta_{0M} \in \beta\} \quad (2.5)$$
3) Starting from the next generation i ($i \geq 1$), a series of cylinders Cyl_i are generated. Cyl_i progress for a length of Δstep compared to the $Cyl_{(i-1)M}$. Cyl_i is parameterized as follows:
$$Cyl_i = \{(h_i, r_i, \alpha_i, \beta_i) | h_i = h_0, r_i \in [c_1 r_{(i-1)M}, c_2 r_{(i-1)M}], \alpha_i \in \alpha, \beta_i \in \beta\} \quad (2.6)$$
The same procedure is used to select Cyl_{iM} (see Figure 2.3);
4) Terminate tracking at iteration i where
$$i = \{r_{iM} > c_3 r_{kM} \cup r_{iM} < c_4 r_{kM} | k \in [i-10, i-1]\} \quad (2.7)$$



(a) (b) (c)

Figure 2.4: Aorta surface refinement. (a) a section of the triangular mesh model with unit diameter; (b) centerline P_i, center points x_i and the perpendicular image rays R_i; (c) the detected aorta surface represented by the triangular mesh.



(a) (b) (c)

Figure 2.5: Ray averaging model and final segmentation. (a) ray averaging model; (b) final segmented aorta (red) overlaid on the intensity image in an axial slice; (c) final segmented aorta (red) together with lungs (pink) and bones (light grey) in a sagittal view.

Algorithm 2.3 Ray casting algorithm

1) Denote the center points of the best matching cylinders as P={P_0, P_1, …P_N}. Partition P into k continuous subsets with approximately the same size. In each subset, fit a 3rd order polynomial function to the points to form a polynomial centerline CL_i;
2) Map a triangular mesh model with unit diameter to CL_i, i∈[1, 2,…, k]. The mesh model interval between layers is L mm (see Figure 2.4 (a));
3) Cast image rays perpendicular to CL_i through the center of each triangle (see Figure 2.4 (b)). Denote each ray as R and the sampled intensity values along R as r_1, r_2, …, r_t. Denote the 12 neighboring rays of R as N1 to N12 and the sampled intensity along Ni as r_{1Ni}, r_{2Ni}, …, r_{tNi}, compute the averaged intensity for R (see Figure 2.5 (a)):
$$\bar{r}_i = \frac{r_i + \sum_{k=1}^{k=12} r_{iNk}}{13} \quad (2.8)$$
4) Based on the averaged values, determine a termination location for each R based on a pre-set fat tissue threshold T_f, where
$$\underset{i}{\operatorname{argmin}}\, i = \{\bar{r}_i < T_f | i \in [1, \dots, t]\} \quad (2.9)$$
5) If no such termination location can be found, terminate R when it first intersects with the a labeled pixel.
6) Together the termination points are used to construct the final aorta surface (see Figure 2.5 (b)-(c)).

Aorta segmentation parameters are listed in Table 2.1. Parameters used in cylinder tracking are crucial to the segmentation success. Several aspects are taken into considerations when designing the tracking parameters:

1) Tracking parameters should be relaxed enough so that tracking does not terminate too early, i.e. a small percentage of labeled pixels should be allowed in the cylinder regions and a small

diameter variation should be allowed;

2) Ascending tracking parameters should be relaxed enough so that the cylinders can go around the curved aortic arch;

3) Tracking parameters should be strict enough so that the model cylinders do not turn into other unlabeled regions outside aorta;

Table 2.1. Aorta segmentation parameters.

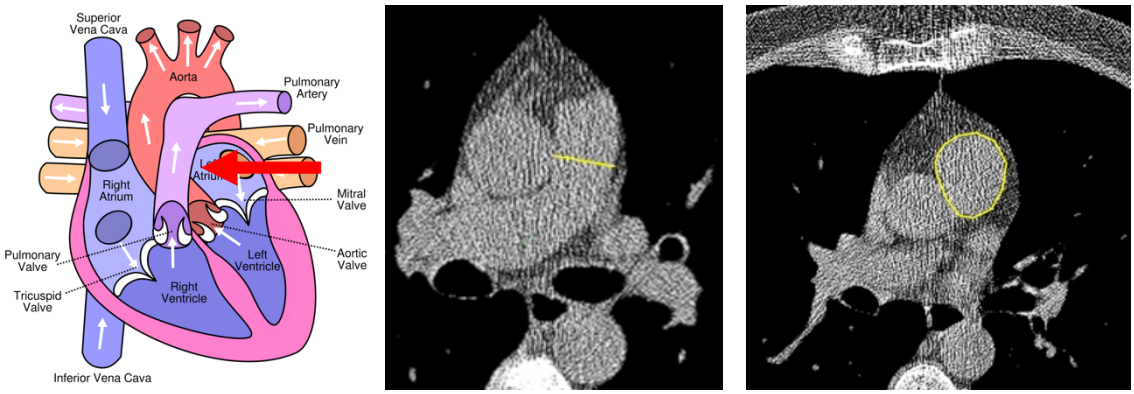
Parameter	Meaning	Value
h_0	Model cylinder height	20mm
r_0	Initial cylinder radii range	[8mm, 30mm]
α	Cylinder azimuthal angle	[0, 0.2π] for descending; [0, 0.25π] for ascending;
β	Cylinder polar angle	[0, 2π]
ω	Coefficient in the matching score function	30
c_1 and c_2	Radii change percentage	c_1=0.5 and c_2=1.5 for ascending; c_1=0.8 and c_2=1.2 for descending;
c_3 and c_4	Radii change percentage upper and lower limits	c_3=2 and c_4=0.5
Δstep	Cylinder progression length	4mm
L	Ring interval in the triangular mesh model	4mm
T_f	High threshold for fat pixels	-30HU

2.1.2 Segmentation of the pulmonary artery trunk

Pulmonary artery carries deoxygenated blood from the heart to the lungs. It originates from the right ventricle and has an almost vertical section (PT) before branching into the two lungs (see Figure 2.6 (a) [53]). An enlarged pulmonary artery diameter is associated with pulmonary hypertension (Edwards et al. 1998, Devaraj et al. 2010) [54-55]. Conventional measurement of pulmonary artery diameter is usually performed manually on the main pulmonary artery at the level of bifurcation (see Figure 2.6 (b)). However, in low-dose non-contrast CT, pulmonary artery in this region does not usually have a well-defined shape or boundary and is difficult to be robustly segmented.

An alternative segmentation and measurement is used in this study: the pulmonary artery

trunk region (see Figure 2.6 (a), (c)) adjacent to the AA is segmented and measured instead of the main pulmonary artery. Compared to main pulmonary artery, this region is better defined and constrained by aorta and fatty tissues and can be more robustly segmented.



(a) (b) (c)

Figure 2.6: Example images of pulmonary artery. (a) a diagram showing the pulmonary artery with the red arrow pointing to PT; (b) the conventional manual measurement of main pulmonary artery diameter; (c) proposed segmentation of PT adjacent to the AA.

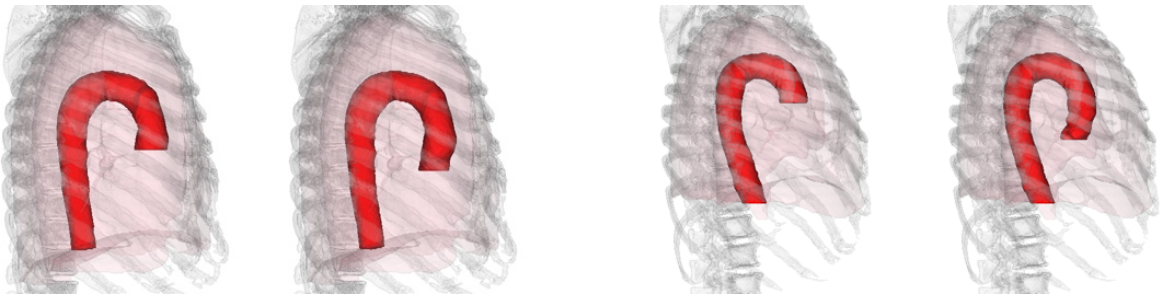
The PT segmentation algorithm (Xie et al. 2015) [56] is listed in Algorithm 2.4. It is performed in the unlabeled region adjacent to the AA. First the segmented AA is extended to ensure that it covers the whole PT region; then PT approximate centerline is determined; finally, a similar ray casting with triangular mesh model is used to detect its surface.

Algorithm 2.4 Pulmonary artery trunk segmentation

1) Extend the segmented AA in a slice-by-slice fashion by projecting it downwards into the unlabeled circular region. In each slice i, denote the unlabeled region as R_i and the major and minor axis of R_i as R_{i1} and R_{i2}, terminate projection in slice i where,
$$\underset{i}{\operatorname{argmin}}\, i = \{R_{i2} < 0.8 * R_{i1}\} \quad (2.10)$$
2) Obtain the new aorta centerline from the projection and re-compute aorta surface as AA_{new} (see Figure 2.7);
3) Detect PT seed point S_{pt} axially at the level of AA_{new} by using a fixed-sized circle of radius s_0 in the unlabeled region adjacent to AA_{new}. This circle C_{min} has a minimized distance to the left and anterior boundaries and S_{pt} is the center of C_{min} (see Figure 2.8 (a)-(b));
4) Use the same triangular mesh and ray casting method described in Algorithm 2.3 to obtain PT surface location (see Figure 2.8 (c)-(d)).

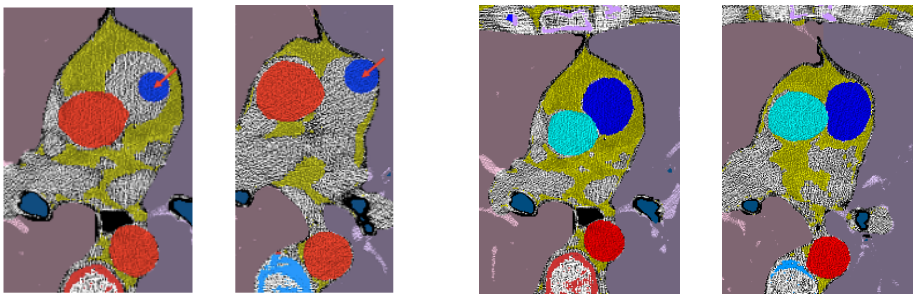
The circle C_{min} has a radius s_0 of 10mm. Compared to aorta segmentation, PT

segmentation is simpler since it is a relatively small region well-bounded by the aorta, the lung and the fat tissues.



(a) (b)

Figure 2.7: Two different cases before and after applying AA extension. For each pair of images, left is before and right is after applying the modification.

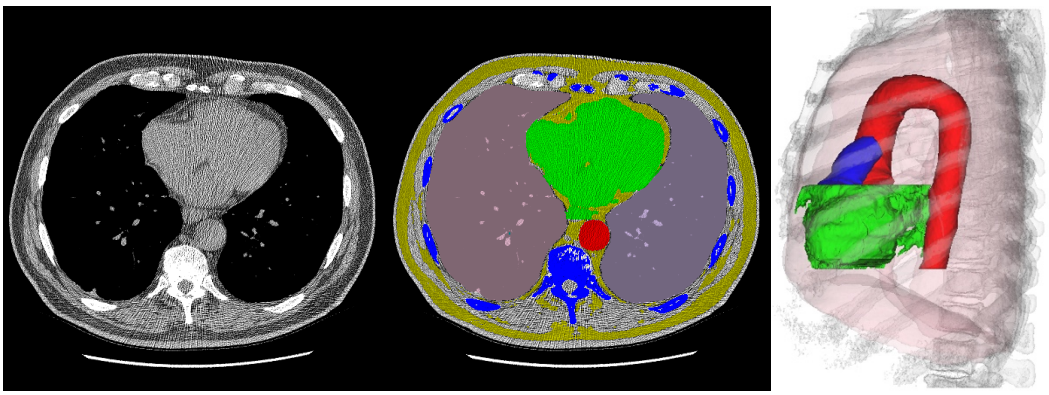


(a) (b) (c) (d)

Figure 2.8: Examples images showing (a)-(b) seed point detection and (c)-(d) final segmentation of PT in two cases. In (a) and (b), aorta is in red and seed point circle is in blue with the red arrows indicating the seed points; in (c) and (d) DA is in red, AA is in light blue and PT is in blue. Other segmented organs such as lungs, fat and vertebrae are also colored.

2.1.3 Segmentation of the heart region

Segmentation of the heart region establishes an important landmark for the segmentation of CAC, one of the most important cardiac biomarkers. However, in LDCT the exact heart boundary is generally not detectable due to the high level of image noise and the lack of contrast (see Figure 2.9). There is only strong evidence for the general heart region, i.e. the thoracic region inferior to the aorta, superior to the abdomen, bounded by the two lungs and fatty tissues. Therefore, the heart region is defined by the constraints from adjacent pre-segmented organs. Heart region segmentation (Xie et al. 2014) [57] algorithm is described in Algorithm 2.5.



(a) (b) (c)

Figure 2.9: Heart region defined by constraints. (a) intensity CT image; (b) heart region (green) constrained by aorta (red), fat (yellow) and lungs (pink); (c) heart region (green) in sagittal view constrained by aorta (red) and PT (blue).

Algorithm 2.5 Heart region segmentation

1) Obtain heart region left and right boundaries Hrt_L and Hrt_R using the segmented lungs;
2) Obtain heart region anterior boundary Hrt_A using segmented fat tissues and sternum;
3) Obtain heart region posterior boundary Hrt_P using the DA;
4) Use AA and DA to define the vertical extent of the heart region Hrt_S and Hrt_I;
5) Determine a unlabeled region I. Denote the left, right, anterior, posterior, superior, and inferior boundaries of I as I_L, I_R, I_A, I_P, I_S, I_I, I is determined as:
$$I = \{I_L > Hrt_L \cap I_R < Hrt_R \cap I_A < Hrt_A \cap I_P > Hrt_P \cap I_S < Hrt_S \cap I_I > Hrt_I\} \quad (2.11)$$
6) Perform connected-component analysis on I to obtain connected regions I_{cc1}, I_{cc2}, …, I_{cck}. Select the largest I_{cci} as the heart region (see Figure 2.9).

2.2 Segmentation and measurement of cardiac biomarkers

Based on the cardiac segmentation, related image biomarkers are measured. In this section, 5 different biomarkers are segmented and/or measured: section 2.2.1-section 2.2.4 correspond to the following biomarkers: CAC content, aorta and PT diameter, AC content, and cardiac fat content. While some of the biomarkers can be easily obtained through intensity thresholding in a mask region, others such as the CAC content require more complex spatial modelling.

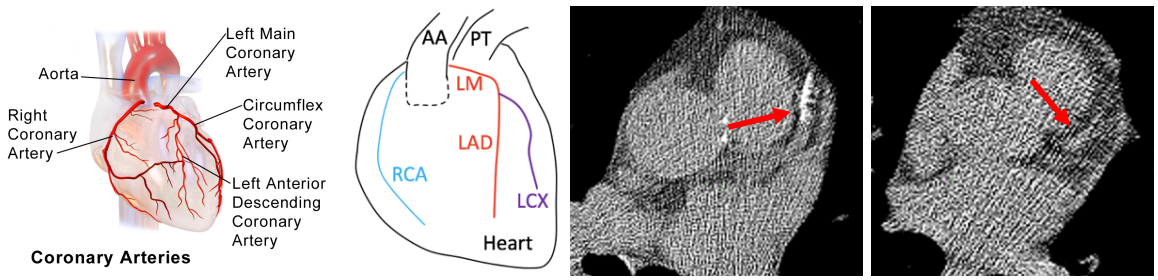
2.2.1 Segmentation and measurement of CAC

CAC is one of the most important biomarkers in the cardiac region. It is an indicator of

coronary heart disease, the most common type of heart disease [4]. In this work, the analysis of CAC consists of two goals: the fully-automated segmentation of all CAC, and the labeling of CAC based on arteries.

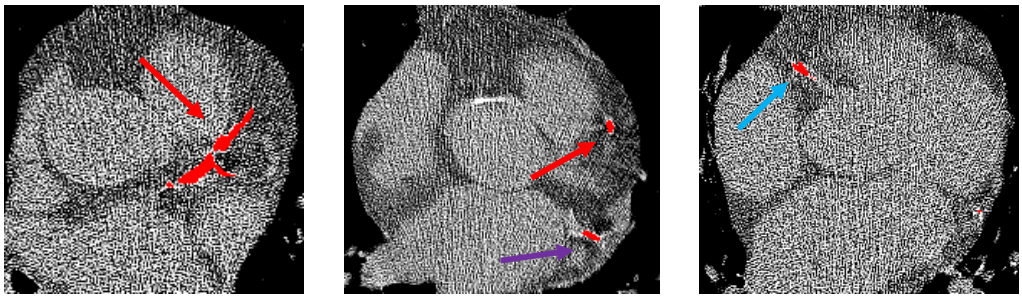
Coronary arteries are located at the heart surface (see Figure 2.10 (a)-(b) [58]). There are four main arteries: Left Main Artery (LM, sometimes also named left coronary artery), Left Anterior Descending (LAD), Left Circumflex (LCX), and Right Coronary Artery (RCA). However, in LDCT they are usually not distinguishable from heart muscles unless they are calcified (see Figure 2.10 (c)-(d)). Therefore, the segmentation of CAC relies on the geometric modeling of calcified regions rather than the explicit segmentation of coronary arteries. Figure 2.11 shows example CACs in different arteries. The CAC segmentation algorithm is described in Algorithm 2.6.

Labeling of CAC by artery location provides the opportunity to evaluate the CAC content in each individual artery. Tota-Maharaj et al. (2015) [59] demonstrated that LM and LAD CAC had independent prognostic significance in predicting mortality by manually labeling and scoring CAC content by the 4 main artery groups (LM, LAD, LCX, RCA). In this dissertation, the automated labeling combined LM and LAD CAC into one label due to the difficulty to visually distinguish them at the border, which is similar to other automated CAC labeling algorithms such as that by Wolterink et al. (2015) [19] and Ding et al. (2015) [60]. The algorithm produces 3 labels: LM+LAD CAC, LCX CAC, RCA CAC. The labeling algorithm is listed in Algorithm 2.7 (Xie et al. 2017) [61].



(a) (b) (c) (d)

Figure 2.10: Diagrams and examples of coronary arteries. (a) diagram of the four main coronary arteries within the heart; (b) simplified diagram depicting the geometric locations of the coronary arteries relative to the heart, AA and PT; (c) a case where the coronary artery is visible because of CAC; (d) a case where the coronary artery is indistinguishable due to the lack of CAC.

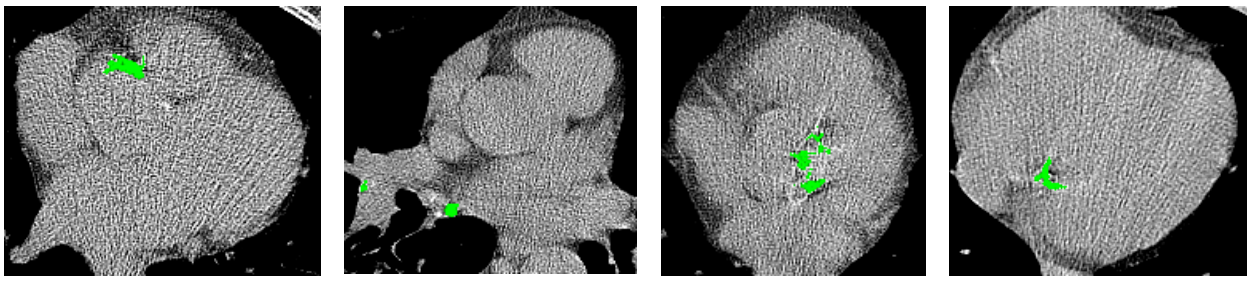


(a) (b) (c)

Figure 2.11: Example CACs in different arteries. (a) LM+LAD CAC (red arrow); (b) LAD CAC (red arrow) and LCX CAC (purple arrow); (c) RCA CAC (blue arrow).

Algorithm 2.6 Segmentation of CAC

1) Pre-filter the intensity CT image I with filter S_f. The size of S_f varies based on the heart region noise level Hrt_N (pixel standard deviation). Denote the filtered image as I_f;
$$I_f = I \otimes S_f \quad (2.12)$$
2) Denote the segmented heart region as Hrt_s, dilate it to Hrt_D with a kernel k_D to include all the potential CAC candidates;
$$Hrt_D = Hrt_s \oplus k_D \quad (2.13)$$
3) Threshold I_f in Hrt_D at T_c to obtain all CAC candidates I_{cac1}. Denote a pixel in I_f as I_p,
$$I_{cac1} = \{I_p \geq T_c | I_p \in Hrt_D\} \quad (2.14)$$
4) Exclude I_p on the surface of aorta (AC, see Figure 2.12 (a)) or adjacent to AA root (aortic valve calcification, see Figure 2.12 (c));
5) Exclude candidates in the lung heart transition region (artifacts, see Figure 2.11 (b)) and candidates in the posterior central heart region (mitral valve calcification, see Figure 2.11 (d)).



(a) (b) (c) (d)

Figure 2.12: Examples of excluded high-intensity regions. (a) AA calcification; (b) calcium and artifacts in lung-heart transition region; (c) aortic valve calcification; (d) mitral valve calcification.

The parameters in Algorithm 2.6 is listed in Table 2.2. In an ideal situation (low image noise), no filtering is needed and the standard calcification threshold of 130HU is sufficient. However, in LDCT scans, no filtering and standard threshold would result in noise pixels being mistaken as calcification. On the other hand, a larger filter window or higher threshold value may lead to the loss of small regions of calcification. Therefore, filter sizes and threshold values are designed to reduce the effect of noise while minimizing the loss of calcification.

Table 2.2. CAC segmentation parameters.

Parameter	Meaning	Value
S_f	Mean filter window size	3 pixels for Hrt_N<=150HU; 5 pixels for Hrt_N>150HU;
T_c	Calcification threshold	130HU for superior heart; 180HU for inferior heart;
k_D	Dilation kernel size	30 mm

CAC labeling relies on the spatial constraints from pre-segmented organs such as the aorta, heart region and pulmonary artery trunk. The general labeling model is shown in Figure 2.12. Labeling starts from the most easily identified CAC regions (RCA and LM+LAD CAC) and use them as constraints for the more subtle CAC regions. The algorithm details are listed in Algorithm 2.7.

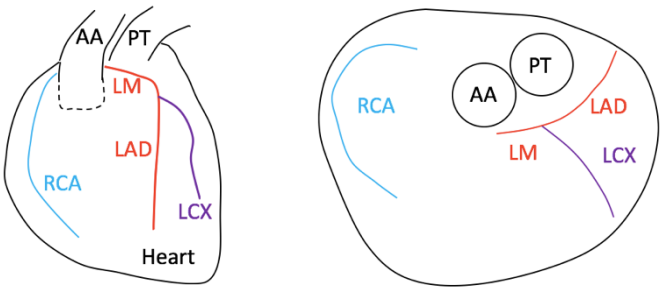


Figure 2.13: CAC labeling model. The four main arteries are shown together with the heart region, AA, and PT.

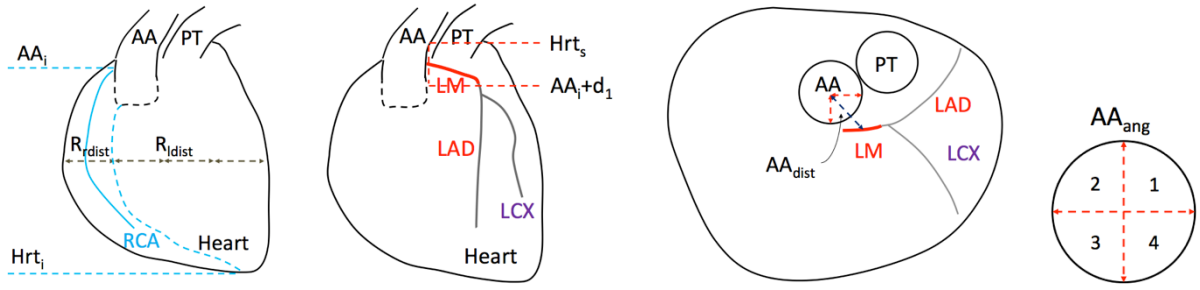
Algorithm 2.7 Labeling of CAC

1) Denote all CAC candidates as I_{CAC}. Denote the segmented left and right lung as I_{LLung} and I_{RLung}. Denote the inferior AA margin as AA_i and the inferior heart margin as Hrt_i. Segment RCA CAC I_{RCA} by applying constraints (see Figure 2.14 (a)):
$$R_{ldist} = distance(I_{LLung}, I_p), \quad I_p \in I_{CAC} \tag{2.15}$$
$$R_{rdist} = distance(I_{RLung}, I_p), \quad I_p \in I_{CAC} \tag{2.16}$$
$$I_{RCA} = \{R_{ldist} > \omega_R * R_{rdist} | I_p \in I_{CAC}, I_p > AA_i \cap I_p < Hrt_i\} \tag{2.17}$$
where ω_R is a weight coefficient;
2) Segment LM+LAD CAC adjacent to aorta I_{LMD1}. Denote the relative location of I_{CAC} to AA as AA_{ang} (shown in Figure 2.14 (d)). Denote the superior heart margin as Hrt_s.
$$R_{adist} = distance(AA, I_p), \quad I_p \in I_{CAC} \tag{2.18}$$
$$I_{LMD1} = \{R_{ldist} < \omega_L * R_{rdist} \cap R_{adist} < AA_{dist} \cap AA_{ang} = 4 | I_p \in I_{CAC}, I_p < AA_i + d_1 \cap I_p > Hrt_s\} \tag{2.19}$$
where AA_{dist} and d_1 are distance margins and ω_L is a weight coefficient (see Figure 2.14 (b)-(d));
3) Label remaining left CACs. If any I_{LMD1} is detected in step 2, the remaining CACs anterior to it is labeled as LM+LAD CAC and those posterior to it is labeled as LCX CAC; otherwise proceed to step 4 (see Figure 2.15 (a));
4) Label the left CACs adjacent to PT as I_{LMD2} and perform anterior/posterior partitions based on I_{LMD2} as shown in Figure 2.15 (b). If no I_{LMD2} is detected, label I_{LMD3} based on its location to AA and performs partitions using I_{LMD3} as shown in Figure 2.15 (c).
$$I_{LMD2} = \{distance(PT, I_p) < PT_{dist} | I_p \in I_{CAC}\} \tag{2.20}$$
$$I_{LMD3} = \{AA_{ang} = 1 | I_p \in I_{CAC}\} \tag{2.21}$$

The parameters used in CAC labeling is listed in Table 2.3. The parameters are designed to be general enough to accommodate for anatomical variations within the heart region. At the same time, they should also be strict enough to not mistake one region for another.

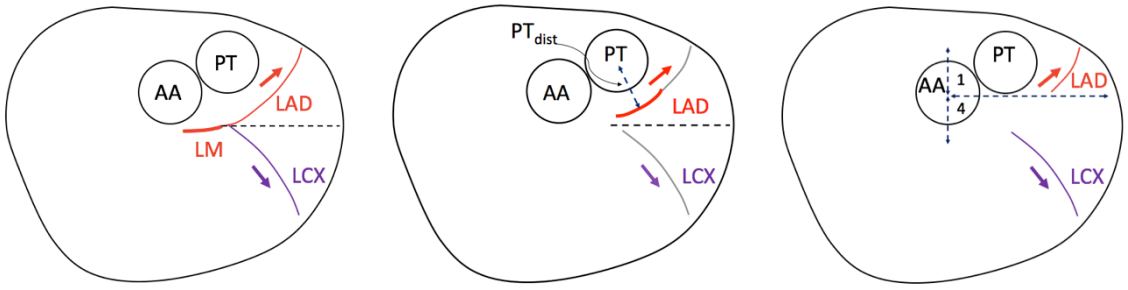
Table 2.3. CAC labeling parameters.

Parameter	Meaning	Value
ω_R	weight coefficient for RCA CAC distance to the lungs	3
AA_{ang}	LM+LAD CAC location relative to AA	4
AA_{dist}	Maximal distance between LM+LAD CAC and AA	30mm
d_1	Maximal distance between LM+LAD CAC and inferior aorta	15mm
ω_L	weight coefficient for LM+LAD CAC distance to the lungs	1
PT_{dist}	Maximal distance between LM+LAD CAC and PT	40mm



(a) (b) (c) (d)

Figure 2.14: Labeling of RCA CACs and LM+LAD CACs adjacent to aorta. (a) RCA CAC labeling; (b) and (c) show the labeling of LM+LAD CACs adjacent to aorta; (d) AA relative angle computation.



(a) (b) (c)

Figure 2.15: Labeling of remaining left region CACs. (a) partition model when there is calcification adjacent to AA; (b) partition model using CAC adjacent to PT; (c) partition model using the AA.

2.2.2 Measurement of aorta diameter profile and PT diameter

Enlarged aorta diameter may be an indicator of aorta aneurysms and impending ruptures (Schwartz et al. 2007) [62], therefore making the early detection of abnormal aorta diameter very important. Pulmonary artery diameter has been traditionally measured in an axial slice at the bifurcation level to assess for pulmonary hypertension risk. Sometimes pulmonary artery

diameter to AA diameter ratio is measured as an alternative to pulmonary artery diameter (Edwards et al. 1998, Devaraj et al. 2010, Karazincir et el. 2008, Truong et al. 2012) [54-55, 63-64]. In our study, both PT and AA equivalent diameters are measured. The aorta diameter profile is also computed. The measurement algorithm is summarized in Algorithm 2.8.

Algorithm 2.8 Measurement of aorta and PT diameter

1) Measure equivalent aorta diameter from triangular mesh model (see Algorithm 2.3). For each ring of triangles perpendicular to the centerline, denote its cross-sectional area as A_i. The equivalent diameter for each ring is denoted as d_i:

$$d_i = 2\sqrt{\frac{A_i}{\pi}} \tag{2.22}$$

Starting from AA to DA, the equivalent diameter for each ring $\{d_1, d_2, \ldots, d_N\}$ is the aorta diameter profile (see Figure 2.16 (a));

2) Obtain the set of equivalent diameters of the AA $\{d_{AA1}, d_{AA2}, \ldots, d_{AAm}\}$ and compute its average:

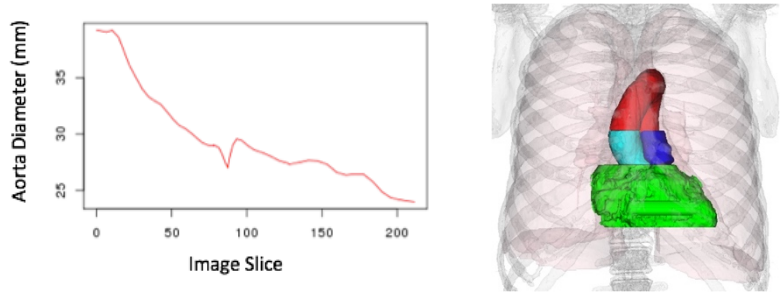
$$\bar{d}_{AA} = \frac{1}{m}\sum_{i=1}^{i=m} d_{AAi} \tag{2.23}$$

3) Similarly obtain the set of equivalent diameters in the PT triangular mesh model $\{d_{PT1}, d_{PT2}, \ldots, d_{PTn}\}$ and compute its average:

$$\bar{d}_{PT} = \frac{1}{n}\sum_{i=1}^{i=n} d_{PTi} \tag{2.24}$$

4) Compute the diameter ratio d_R as shown in Figure 2.16 (b).

$$d_R = \frac{\bar{d}_{AA}}{\bar{d}_{PT}} \tag{2.25}$$



(a) (b)

Figure 2.16: Aorta and PT diameter measurement. (a) aorta diameter profile from ascending to descending; (b) region where the diameters are measured (light and dark blue). Heart (green) and lungs (pink) are also shown.

2.2.3 Segmentation and measurement of AC

Thoracic AC is associated with an increased risk of cardiovascular disease and death (Eisen et al. 2008) [65]. It manifests in a CT image to be high-intensity calcium deposits on the

aorta surface (see Figure 2.17 (a)-(b)). Using the traditional Agatston scoring method (Agatston et al. 1990) [66] used for CAC scoring, AC tends to have a much higher Agatston score (AS) compared to CAC (see Figure 2.17 (c)-(e)).

(a) (b) (c) (d) (e)

Figure 2.17: Examples of AC. (a) calcification in AA and DA (red arrows); (b) aortic arch calcification (red arrows); (c)-(e) three cases with different levels of AC. The AS for (c), (d) and (e) is respectively 302, 3311 and 9088.

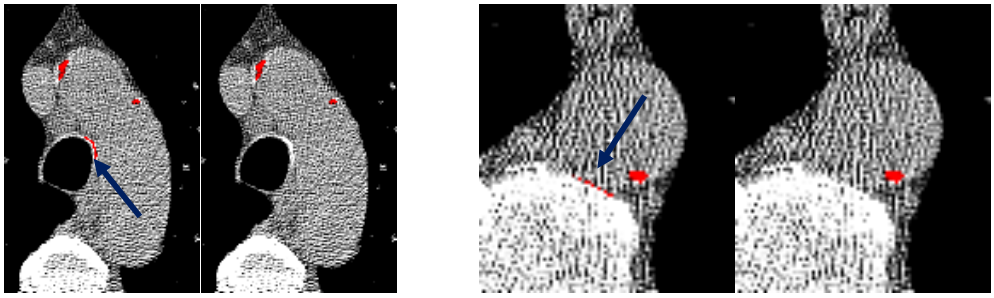
AC can be measured in the aorta mask region (Xie et al. 2014) [67] using the method listed in Algorithm 2.9. The algorithm is similar to that used for the CAC segmentation (Algorithm 2.6). Compared to CAC, AC is less likely to be confused with other calcification since it is only located on the surface of aorta. Therefore, a simpler method is used. The associated parameters are listed in Table 2.4. Similar to CAC segmentation parameters, the filter size and threshold value are selected to minimize the impact of noise while retaining the largest amount of detectable calcification.

Algorithm 2.9 Segmentation of AC

1) Pre-filter the intensity CT image I with a filter of size S_A to reduce the effect of noise;
$$I_f = I \otimes S_A \quad (2.26)$$
2) Dilate the segmented aorta A_s with a kernel k_D to A_D to include all potential calcification;
$$A_D = A_s \oplus k_D \quad (2.27)$$
3) Extract high intensity candidates I_{AC1} in A_D by thresholding at T_A;
$$I_{AC1} = \{I_p \geq T_A | I_p \in A_D\} \quad (2.28)$$
4) Perform connected component analysis and eliminate candidate smaller than V_A pixels as noise regions. The refined candidates are denoted as I_{AC2}:
$$I_{AC2} = \{volume(I_p) \geq V_A | I_p \in I_{AC1}\} \quad (2.29)$$
5) Eliminate candidates adjacent to segmented bone and airway surface as bone or airway calcification or airway artifacts (see Figure 2.18 for exclusion examples).

Table 2.4. AC segmentation parameters.

Parameter	Meaning	Value
S_A	Mean filter window size	3 pixels
T_A	Calcification threshold	160HU
V_A	Minimal calcification volume	5 pixels
k_D	Dilation kernel size	5mm



(a) (b)

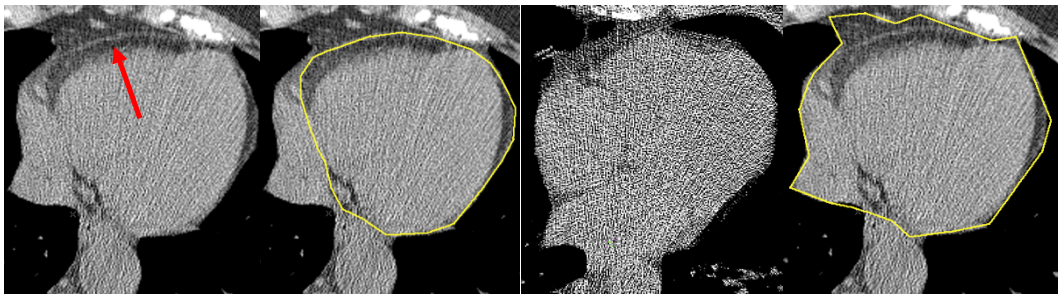
Figure 2.18: Exclusion of non-AC artifacts. (a) shows airway calcification (blue arrow); (b) shows bone pixels (blue arrow). All high-intensity regions are colored red. For each pair, left is before exclusion and right is after exclusion.

2.2.4 Segmentation and measurement of cardiac visceral fat

Fat around the heart region is associated with cardiovascular diseases (Rosito et al. 2008) [68]. The most commonly measured cardiac fat is pericardial fat, i.e. fat enclosed by the visceral pericardium (see Figure 2.19 (a)-(b)). However, in LDCT the pericardium region is not always clearly visible (see Figure 2.19 (c)). Therefore, an alternative approach is adopted to measure all

the fat in the cardiac visceral region – cardiac visceral fat (CVF). It is defined to be at the same level of the segmented heart, enclosed by lungs and posterior to the sternum (see Figure 2.19 (d)).

Similar to heart region segmentation, CVF segmentation relies on the constraints from adjacent pre-segmented organs. However, the anterior region needs further refinement to stop the CVF from leaking out of the sternum. The algorithm listed in Algorithm 2.10 is used to segment CVF (Xie et al. 2015) [69].

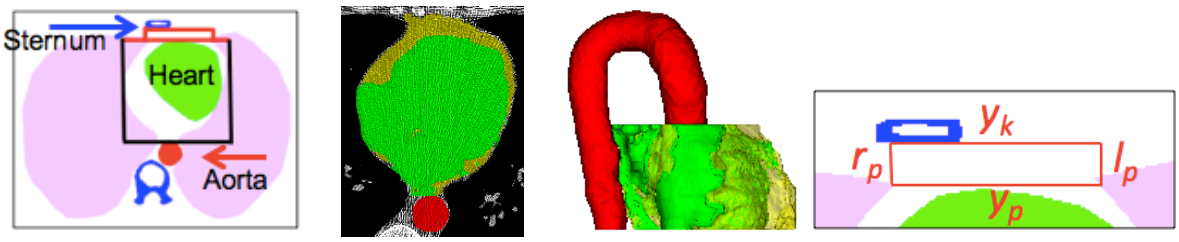


(a) (b) (c) (d)

Figure 2.19: Pericardial fat and CVF. (a) intensity CT image with red arrow pointing to the pericardium; (b) pericardial region (yellow outline) in (a); (c) intensity CT image without visible pericardium; (d) cardiac visceral region (yellow outline) in the same image as (a) and (b).

Algorithm 2.10 Segmentation of cardiac visceral fat

1) At the level of heart, constrain the CVF using the segmented lungs, sternum, DA or vertebrae (see Figure 2.19 (a));
2) Refine the fat region between sternum and lungs axially. In the anterior region, if the lungs form a closed boundary, eliminate all fat between sternum and lungs and proceed to step 5. Otherwise proceed to step 3;
3) Draw a horizontal like y_i defined by the sternum posterior margin. Shift y_i towards posterior until it intersects with the lungs at y_k. The intersection points are denoted as l_k and r_k and their distance d_k:
$$d_k = |l_k - r_k| \quad (2.30)$$
4) Continue shift y_k towards the posterior until it reaches y_p:
$$\underset{y_p}{\operatorname{argmin}}\, d_p = \{d_p | d_p = |l_p - r_p|\} \quad (2.31)$$
Together y_k and y_p define the anterior CVF region;
5) Perform connected-component analysis to all the fat pixels in the constrained region. Retain the fat pixels connected to the heart region to eliminate other non-cardiac fat pixels (see Figure 2.20 (b)-(c)).



(a) (b) (c) (d)

Figure 2.20: CVF segmentation. (a) model with the anterior refinement; (b) and (c) show the segmented CVF (yellow) together with the aorta (red) and the heart (green) in an axial slice and in sagittal view; (d) anterior refinement model where the red box indicates the modeled region between the sternum and lungs.

2.3 Quantitative and qualitative evaluation methods

Cardiac region segmentation and biomarker measurements are evaluated quantitatively and qualitatively. Quantitative evaluation is performed on a set of CT scans with reference manual markings or semi-automatically measured reference biomarkers. Qualitative evaluation is performed through the visual inspection of segmentation outcomes. Both evaluation methods are described in detail in section 2.3.1 and 2.3.2.

2.3.1 Quantitative evaluation experiments

Tables 2.5-2.6 list the segmented regions and biomarkers used in the quantitative evaluation as well as the corresponding evaluation metrics, datasets and references. Definitions and justifications of these evaluation metrics are given in Evaluation 2.1 and 2.2. All CT scans used for evaluation are non-contrast chest CT scans. The majority of them are thin-slice (slice-thickness <= 1.25mm) and low-dose and come from the two public datasets: VIA-ELCAP public dataset [70] and the LIDC public dataset [71]. Example reference regions for segmentation and image biomarkers are demonstrated in Figures 2.21-2.23. In general, quantitative evaluation is performed using standard comparison methods such as evaluating the agreement between automated and reference regions through the Dice Similarity Coefficient (DSC) or computing the Pearson correlation coefficient between automated and reference calcification scores.

The only quantitative evaluation based on visual scores rather than manual markings is the CAC categorical visual scores for each artery (see Table 2.6 CAC row and Figure 2.23). Each scan was assigned a categorical visual CAC score for each of the four main arteries (LM, LAD, LCX, RCA) by a radiologist using the criteria established by Shemesh et al. (2010) [21]. The score ranges from 0 to 3: 0=none, 1=mild, 2=moderate,3=extensive. In this evaluation, the visual scores of LM and LAD are added together for LM+LAD CAC assessment. For total CAC assessment, visual scores from the four arteries are added together.

Table 2.5. Quantitative evaluation metrics for cardiac segmentation.

Region	Metric	Reference	Dataset
Aorta	DSC, B_{pn}, μ_d	Manual aorta boundaries (ascending, descending, arch)	630 boundaries from 60 scans [70-71]
PT	DSC	Manual boundary in a central slice in the segmented region	45 boundaries from 45 LDCT [70]
CVF	DSC	Manual boundaries in 3 evenly spaced slices in the segmented region	135 boundaries from 45 LDCT [70]

Evaluation 2.1: Definition and justification for segmentation evaluation metrics

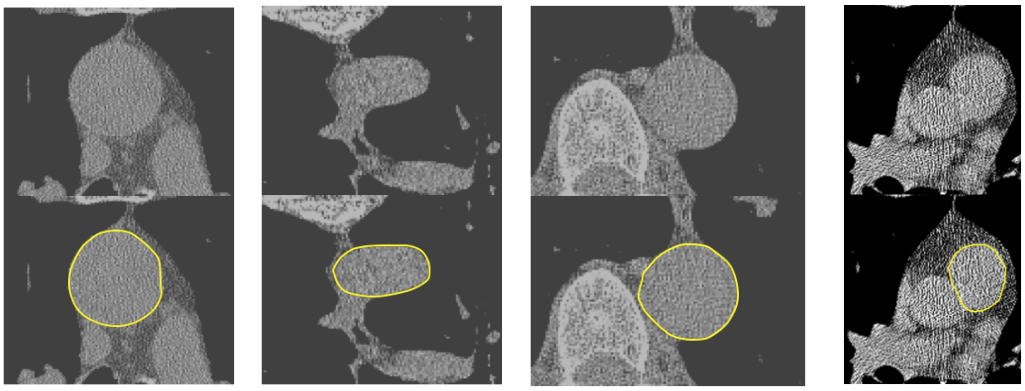
1) Dice Similarity Coefficient (DSC):
$$DSC = 2\left|R_{ref} \cap R_{auto}\right|/(\left|R_{ref}\right| + |R_{auto}|) \quad (2.32)$$
where R_{ref} is the reference region and R_{auto} is the automatically segmented region. DSC is one of the classic and most common region segmentation evaluation metric. If the R_{auto} completely overlaps with R_{ref}, DSC becomes 1; if there is no overlapping, DSC becomes 0. The higher DSC is, the more the two regions agree with each other;
2) Bias term B_{pn}:
$$B_{pn} = \frac{P_{fp} - P_{fn}}{P_{tp}} \quad (2.33)$$
where
$$P_{fp} = \sum I_p, I_p \in R_{auto} \cap \bar{R}_{ref} \quad (2.34)$$
$$P_{fn} = \sum I_p\,, I_p \in \bar{R}_{auto} \cap R_{ref} \quad (2.35)$$
$$P_{tp} = \Sigma I_p\,, I_p \in R_{auto} \cap R_{ref} \quad (2.36)$$
B_{pn} indicates whether the algorithm over-segments ($B_{pn}>0$) or under-segments ($B_{pn}<0$). It provides additional information to DSC since DSC does not distinguish between over- and under-segmentation;
3) A boundary distance d_i is defined as
$$d_i = \underset{d_i}{\text{argmin}}\{\sqrt{(mx_i - ax_i)^2 + (my_i - ay_i)^2 + (mz_i - az_i)^2}\} \quad (2.37)$$
where (mx_i, my_i, mz_i) is a pixel in the boundary of R_{ref} and (ax_j, ay_j, az_j) is a pixel in the boundary of R_{auto}. μ_d is the mean of all d_i. It is more sensitive than DSC and B_{pn} to small boundary changes.

Table 2.6. Quantitative evaluation metrics for cardiac biomarker measurements.

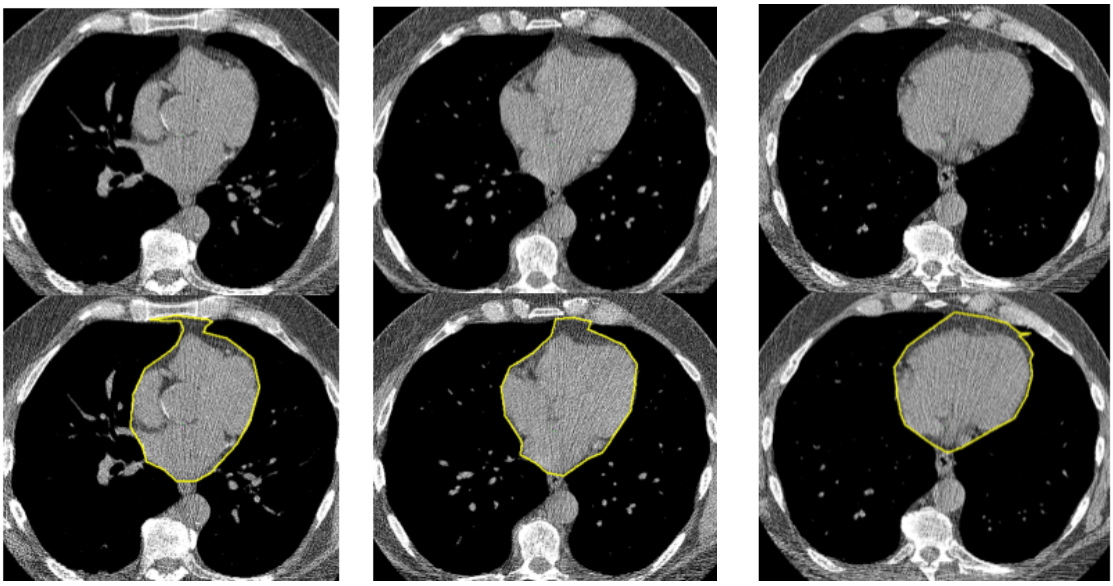
Biomarker	Metric	Reference	Dataset
CAC	Pearson correlation coefficient (AS, Volume Score VS, Mass Score MS), confusion matrix (AS based risk category)	Manual CAC regions	41 LDCT (private cohort)
CAC	Spearman correlation coefficient (AS, VS, MS), labeling sensitivity and specificity (by artery and by noise level)	Visual CAC scores by artery	1,359 LDCT (private cohort)
AC	Pearson correlation coefficient (AS, VS, MS)	Manual AC regions	45 LDCT [70]
PT diameter	Pearson correlation coefficient and Limits of Agreement (LOA)	Manual pulmonary artery diameter at bifurcation	45 LDCT [70]
AA diameter	Pearson correlation coefficient and LOA	Manual AA diameter at bifurcation	45 LDCT [70]
PT/AA ratio	Pearson correlation coefficient and LOA	Manual diameter ratio	45 LDCT [70]

Evaluation 2.2: Definition and justification for biomarker evaluation metrics

1) Agatston-weighted score (AS): clinically defined calcification score. Denote a calcification area as CAC_i and its corresponding weight coefficient as ω_i. Denote the maximum intensity in CAC_i as p_i. AS is defined as:
$$AS = \sum \omega_i * CAC_i \quad (2.38)$$
where
$$\omega_i = \begin{cases} 1, p_i \in (-\infty, 200HU) \\ 2, p_i \in [200HU, 300HU) \\ 3, p_i \in [300HU, 400HU) \\ 4, p_i \in [400HU, +\infty) \end{cases} \quad (2.39)$$
2) Volume score (VS): total volume (mm^3) of the calcification region;
3) Mass score (MS): total volume with each pixel weighted by its HU intensity;
4) Pearson correlation coefficient (Pearson R): ranges from -1 to 1 and characterizes the linear correlation between two continuous variables, where 1 indicates total positive correlation and -1 indicates total negative correlation;
5) Spearman correlation coefficient (Spearman R): ranges from -1 to 1 and characterizes the rank correlation between two continuous or categorical variables;
6) Limits of Agreement 95% (LOA): Bland-Altman 95% LOA [72]. It gives the degree of agreement between method X and method Y. Denote the mean and standard deviation of (X-Y) as μ_{XY} and σ_{XY}, LOA can be approximated as follows:
$$LOA = [\mu_{XY} - 1.96 * \sigma_{XY}, \mu_{XY} + 1.96 * \sigma_{XY}] \quad (2.40)$$
7) Confusion matrix (CM): CM is used to compare the automated risk categories and manual risk categories, which are defined by the automated and manual Agatston scores AS. Risk category = 1 (AS<=10), 2 (AS∈[11, 100]), 3 (AS∈[101, 400]), and 4 (AS>400); CM clearly indicates the disagreement between the two risk categories between different risk levels;
8) Sensitivity: defined by number of true positives TP (both automated and reference CAC scores > 0) and the number of false negatives FN (only reference CAC scores > 0). Sensitivity characterizes the algorithm's ability to capture all the true CAC regions:
$$Sensitivity = TP/(TP + FN) \quad (2.41)$$
9) Specificity: defined by number of true negatives TN (both automated and reference CAC scores = 0) and the number of false positives FP (only automated CAC scores > 0). Specificity characterizes the algorithm's ability to not falsely capture CAC regions:
$$Specificity = TN/(FP + TN) \quad (2.42)$$

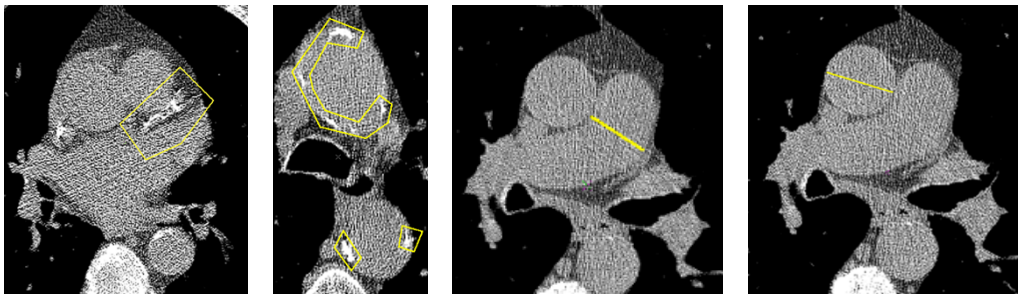


(a) (b) (c) (d)



(e) (f) (g)

Figure 2.21: Example reference regions for quantitative segmentation evaluation. For each pair of images, the upper image is the intensity CT and the lower image shows the manual boundary. (a) AA; (b) aortic arch; (c) DA; (d) PT; (e) superior CVF; (f) middle CVF; (g) inferior CVF.



(a) (b) (c) (d)

Figure 2.22: Example reference regions for quantitative biomarker evaluation. (a) reference region for CAC; (b) reference regions for AC; (c) pulmonary artery reference diameter at bifurcation level; (d) AA diameter at bifurcation level.

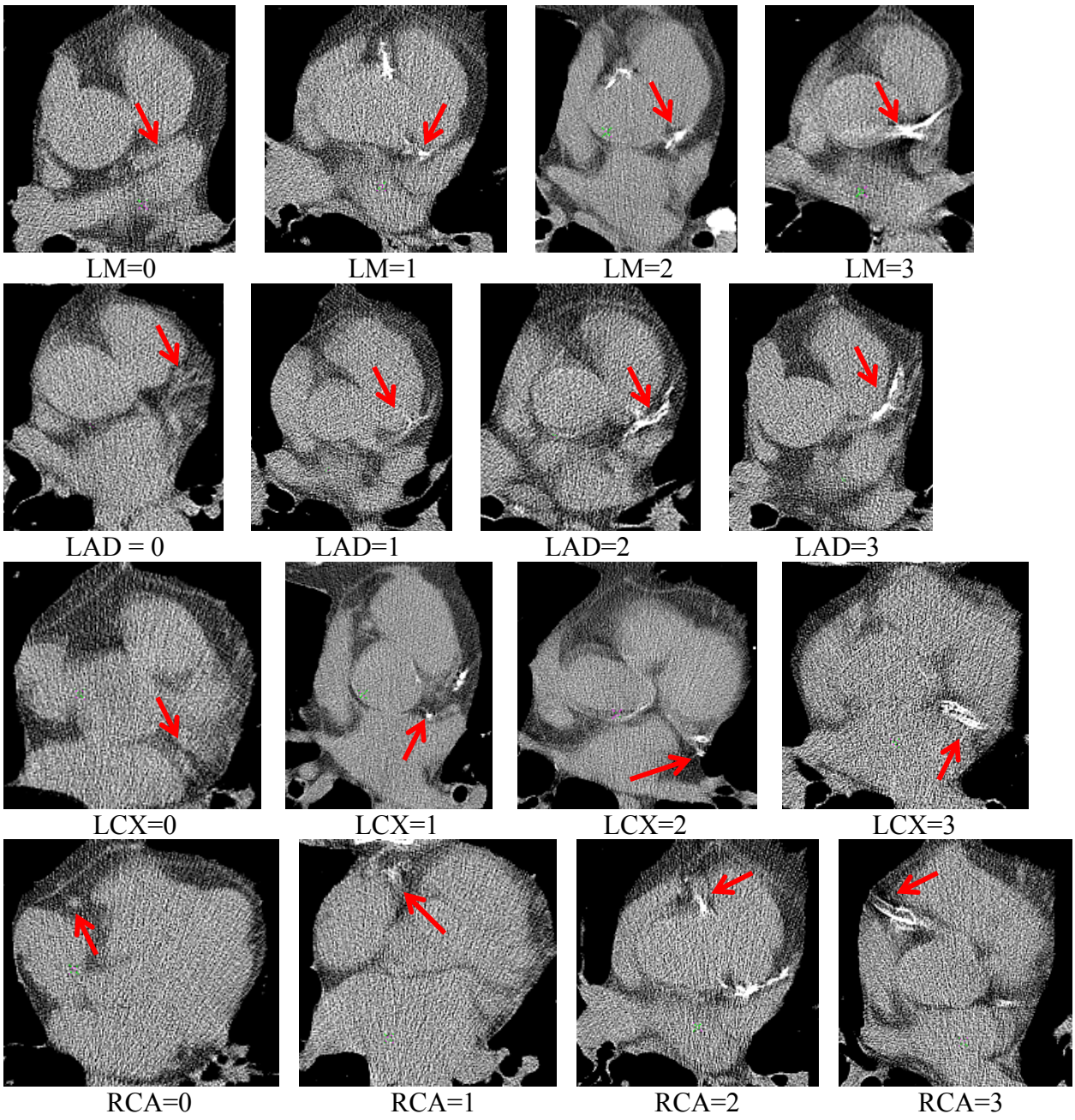


Figure 2.23: Example cases with different visual CAC scores for the four arteries. Red arrows indicate the CAC locations.

2.3.2 Qualitative evaluation experiments

In the above sections, each segmentation algorithm has been quantitatively evaluated in a set of manually marked CT scans (evaluation set typically contains around 50 CT scans). This is also the standard quantitative evaluation procedure used in the literature for segmentation problems. However, to evaluate the general robustness of the algorithm, especially its ability to

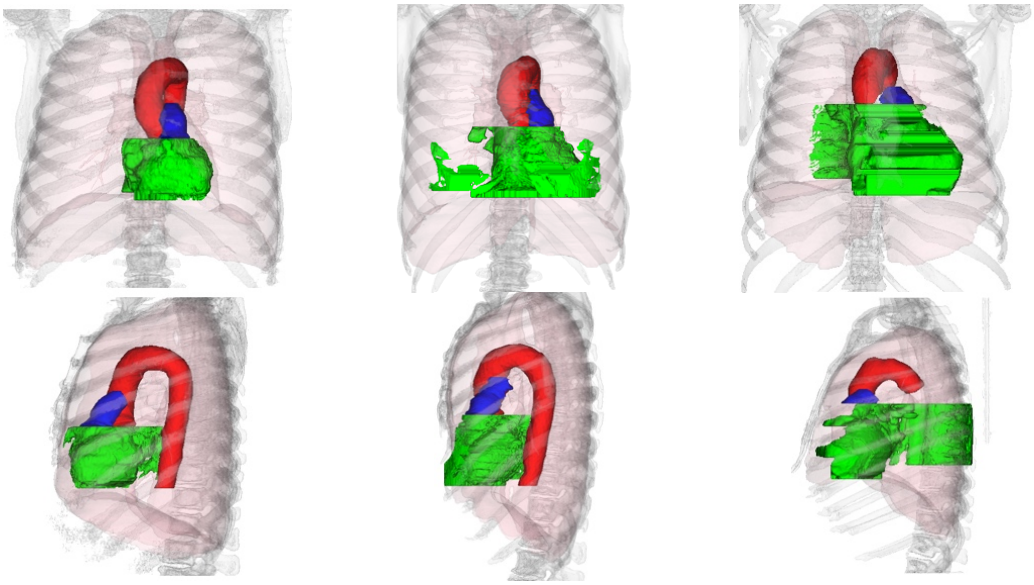
be applied in a fully-automated setting without human intervention, a much larger dataset with thousands of scans is needed. Manually annotating reference regions on such large datasets is typically unrealistic. Therefore, the traditional quantitative evaluation method is not applicable.

In this study, a novel alternative method is used for evaluating the segmentation qualitatively through customized visualization and visual inspection (Reeves et al. 2017) [51]. The visual inspection mainly relies on the coronal and sagittal visualization of the segmented aorta, heart and PT (see Figure 2.24). They are evaluated together as the cardiac region segmentation. Evaluation in 2D axial image slices is also possible (see Figure 2.25) but is rarely used unless there is confusion that cannot be distinguished in the 3D visualization. Each cardiac segmentation is graded into one of the following three categories:

Good: no visible error;

Acceptable: errors exist but the segmentation is still usable for biomarker measurement;

Unacceptable: major errors or algorithm failure, making the segmentation unusable.



(a) (b) (c)

Figure 2.24: Customized visualization of cardiac segmentation in coronal view (upper row) and sagittal view (lower row). The visual segmentation grade is (a) Good; (b) Acceptable; (c) Unacceptable. Aorta (red), heart (green), PT (blue) together with lungs (pink) and bones (light grey) are shown.

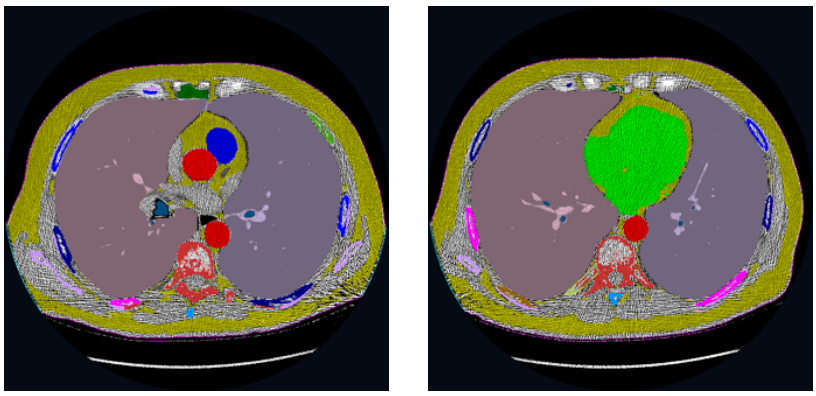


Figure 2.25: Axial visualization of cardiac segmentation in two slices. The segmented aorta (red), heart (green), PT (blue) together with the colored lungs, fat, ribs, vertebrae are overlaid on the intensity CT images.

The datasets for large-scale qualitative evaluation based on visual inspection are: the VIA-ELCAP public dataset [70], the LIDC public dataset [71], a World Trade Center workers dataset WTC (research dataset), a LCS dataset FY from the Flight Attendants Medical Research Institute [73], a nuclear energy workers dataset QV (research dataset), and a high risk LCS dataset UN (research dataset). Number of different cases and imaging parameter settings of these datasets are summarized in Table 2.7. Note that sometimes one case contains multiple scans taken at different time points or taken at the same time but reconstructed with different algorithms. Therefore, the number of cases does not always equal the number of scans. All scans are non-contrast whole-chest CT with the majority being low-dose. Scans with major implant artifacts or very poor quality have been excluded.

Table 2.7. Summary of datasets for large-scale qualitative evaluation of cardiac segmentation.

Dataset	Number of Cases (Include)	Number of Scans (Include)	Number of Scans (Exclude)	Slice Thickness (mm)
VIA-ELCAP	46	46	4 (8%)	=1.25
LIDC	318	318	10 (3%)	<=1.25
WTC	1454	3766	28 (0.7%)	<=1.5 (3762 scans) 2 (4 scans)
FY	932	2137	0 (0%)	<=1.25
QV	1887	4672	6 (0.1%)	=2
UN	2985	9810	10 (0.1%)	<=1.5 (9804 scans) =2 (6 scans)

2.4 Evaluation results

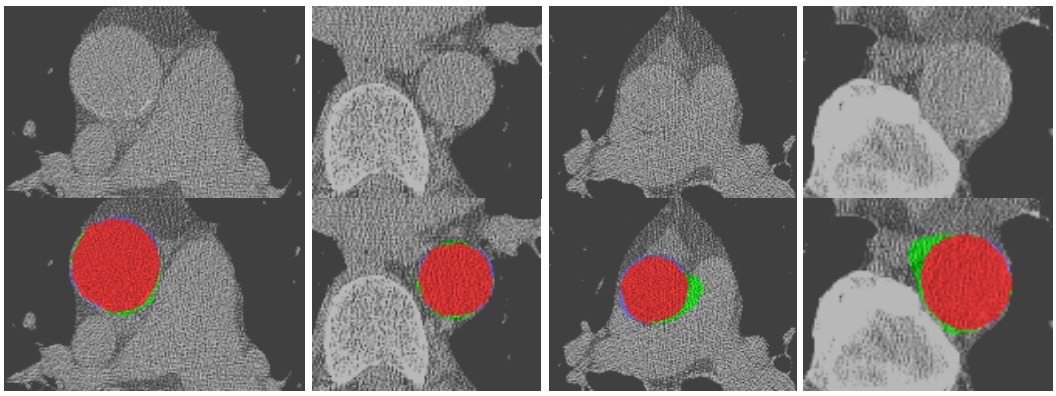
Both quantitative and qualitative evaluation results are given in this section. For each evaluation result, relevant visualizations are also given such as the comparison between automated and manual regions. Section 2.4.1 gives the quantitative evaluation results on segmentation and biomarker measurements and section 2.4.2 gives the large-scale qualitative evaluation results on segmentation.

2.4.1 Quantitative evaluation results

Aorta segmentation results and example images are given in Table 2.8 and Figure 2.26. The average DSC for PT segmentation is 0.88 (range=[0.74, 0.95]). The average DSC for CVF segmentation is 0.93 (range=[0.81, 0.99]). PT and CVF example segmentation outcomes are shown in Figures 2.27-2.28.

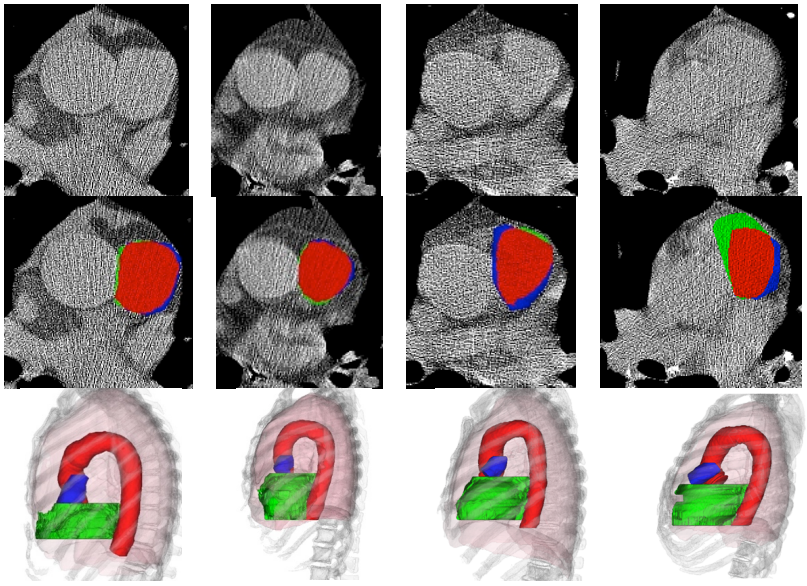
Table 2.8. Quantitative evaluation results for aorta segmentation. Metrics include average DSC for the entire aorta and for AA, DA and arch (Arch); bias term B_{pn}; mean boundary distance μ_d in millimeters for AA, DA and Arch. For each term, the mean, standard deviation (σ), maximum (max) and minimum (min) are given.

	DSC (aorta)	DSC (AA)	DSC (DA)	DSC (Arch)	B_{pn}	μ_d (aorta)	μ_d (AA)	μ_d (DA)	μ_d (Arch)
Mean	0.933	0.936	0.939	0.915	0.018	1.39	1.63	1.20	1.42
σ	0.014	0.019	0.020	0.022	0.054	0.19	0.33	0.29	0.22
Max	0.963	0.972	0.973	0.955	0.123	1.79	2.68	1.96	1.95
Min	0.907	0.889	0.881	0.863	-0.086	0.83	0.69	0.57	0.98



(a) (b) (c) (d)

Figure 2.26: Example aorta segmentation compared to the references in 4 cases. Upper row shows intensity CT image and lower row shows the comparison. In each color image, red is the overlapping region between automated and reference image, green only exists in the automated image and blue only exists in the reference image. The DSC and μ_d for each case are: (a) 0.962 and 1.23mm; (b) 0.957 and 1.08mm; (c) 0.888 and 1.89mm; (d) 0.923 and 1.44mm.



(a) (b) (c) (d)

Figure 2.27: Example PT segmentation compared to the references in 4 cases. Top row shows intensity CT image, middle row shows the comparison, and bottom row shows sagittal visualization of PT (blue), aorta (red) and heart (green). In each color image in the middle row, red is the overlapping region between automated and reference image, green only exists in the automated image and blue only exists in the reference image. The DSC for each case is: (a) 0.94; (b) 0.93; (c) 0.88; (d) 0.78.

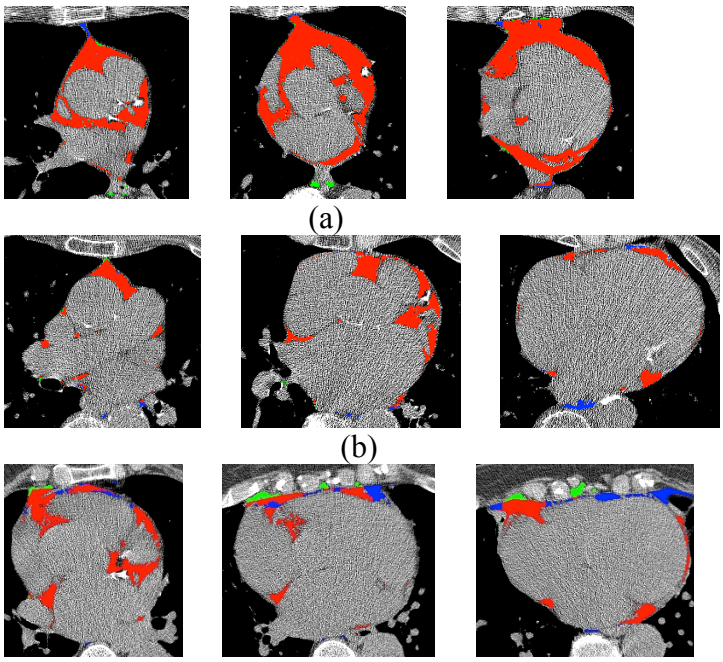


Figure 2.28: Example CVF segmentation compared to the references in 3 cases. Each row corresponds to a case, where the 3 images are the 3 slices with manual boundaries from superior to inferior. For all images, red represents fat pixels in both the automated and reference regions, green represents those only in the automated region and blue only in the reference region.

The quantitative evaluation results for CAC biomarkers are shown in Tables 2.9-2.13. The CAC evaluation results compared to manual boundaries are summarized in Tables 2.9-2.10. The automated calcification score distributions of the CAC datasets used for visual scoring is listed in Table 2.11. The CAC evaluation results compared to visual scores are summarized in Tables 2.12-2.13. After evaluation using the whole dataset, the dataset is further partitioned into three subsets with different noise levels to evaluate the impact of image noise on the algorithm performance. Noise level is represented by the standard deviation of pixels in the automatically segmented heart region. Figure 2.29 shows the heart region noise distribution for the whole dataset. The three subsets are: low noise group (noise below 80HU) with 398 cases; medium noise group (noise between 80 and 110HU) with 499 cases; high noise group (noise above 110HU) with 462 cases. The partition thresholds have been chosen so that each noise group would contain similar number of cases. Results on the three noise groups are shown in Tables

2.14-2.15 and Figure 2.30. Sensitivity differences between the three noise groups are evaluated using Pearson Chi-squared test for its significance. The following sensitivities are significantly different from each other, i.e. p-value<0.05: all CACs low v.s. high (p=0.008); all CACs medium v.s. high (p=0.009); LM+LAD CAC low v.s. high (p<0.001); LM+LAD CAC medium v.s. high (p<0.001). Example labeled CACs are shown in Figure 2.31.

Table 2.9. Pearson R between automated (Auto) and reference scores (Manual) for AS, VS, and MS.

Pearson R	AS	VS	MS
Manual v.s. Auto	0.90	0.89	0.92

Table 2.10. Confusion matrix (CM) between automated (Auto) and reference (Ref) risk categories. Low risk=1; moderate risk=2; moderately high risk=3; high risk=4.

CM	Auto=1	Auto=2	Auto=3	Auto=4
Ref=1	18	1	1	0
Ref=2	1	5	1	0
Ref=3	0	1	5	3
Ref=4	0	0	0	5

Table 2.11. Statistics of AS scores for the whole dataset (all CACs and by artery).

AS	Number of Cases	All CACs	LM+LAD CAC	LCX CAC	RCA CAC
Mean	1,359	391.9	213.5	78.76	99.7
Median	1,359	71.0	35.0	0	0
Range	1,359	[0, 6328.0]	[0, 3657.0]	[0, 4105.0]	[0, 4836.0]
Standard deviation	1,359	762.5	428.1	321.8	321.9

Table 2.12. Spearman R between visual and automated scores (all CACs and by artery).

Spearman R	Number of Cases	All CACs	LM+LAD CAC	LCX CAC	RCA CAC
Visual v.s. AS	1,359	0.87	0.82	0.66	0.72
Visual v.s. VS	1,359	0.88	0.82	0.66	0.73
Visual v.s. MS	1,359	0.87	0.82	0.66	0.72

Table 2.13. Labeling sensitivity and specificity (all CACs and by artery).

Metric	Number of Cases	All CACs	LM+LAD CAC	LCX CAC	RCA CAC
Sensitivity	1,359	0.87	0.82	0.65	0.74
Specificity	1,359	0.84	0.89	0.92	0.88

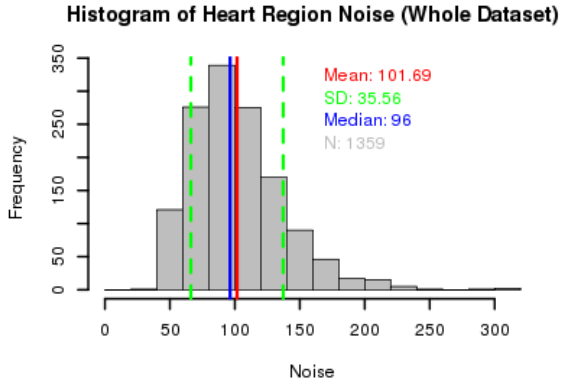


Figure 2.29: Histogram of heart region noise on the whole dataset (1,359 scans).

Table 2.14. Spearman R between visual and automated scores for the three noise groups (all CACs and by artery). 1 = low noise group (398 cases); 2 = medium noise group (499 cases); 3 = high noise group (462 cases).

Spearman R	All CACs			LM+LAD CAC			LCX CAC			RCA CAC		
	1	2	3	1	2	3	1	2	3	1	2	3
Visual v.s. AS	0.87	0.90	0.85	0.83	0.85	0.80	0.68	0.65	0.67	0.74	0.74	0.69
Visual v.s. VS	0.88	0.90	0.86	0.83	0.85	0.80	0.68	0.65	0.68	0.75	0.74	0.69
Visual v.s. MS	0.87	0.89	0.85	0.83	0.85	0.80	0.68	0.65	0.67	0.74	0.73	0.69

Table 2.15. Labeling sensitivity and specificity for the three noise groups (all CACs and by artery). 1 = low noise group (398 cases); 2 = medium noise group (499 cases); 3 = high noise group (462 cases).

Metric	All CACs			LM+LAD CAC			LCX CAC			RCA CAC		
	1	2	3	1	2	3	1	2	3	1	2	3
Sensitivity	0.90	0.89	0.82	0.87	0.85	0.74	0.66	0.66	0.63	0.74	0.78	0.70
Specificity	0.84	0.86	0.83	0.84	0.91	0.91	0.94	0.90	0.92	0.90	0.87	0.88

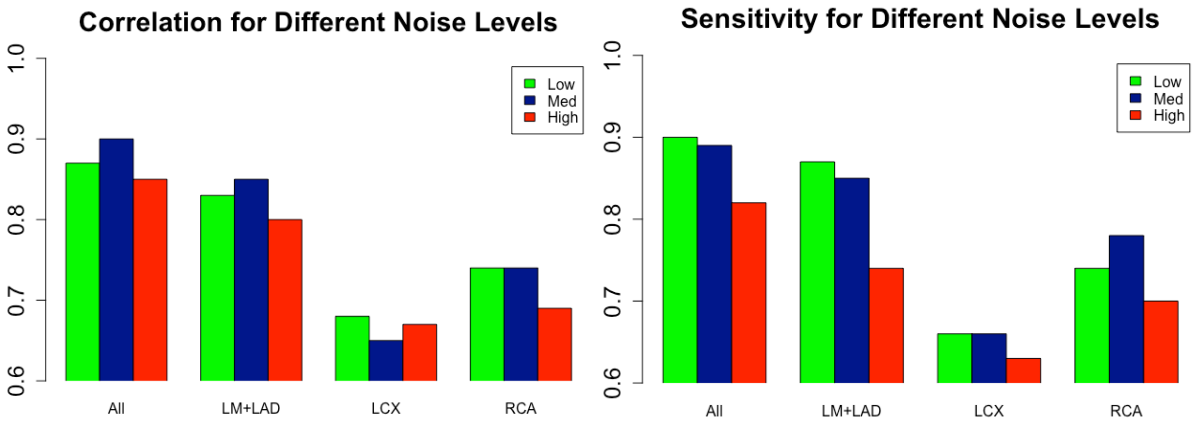


Figure 2.30: Bar plots showing the differences in correlation and sensitivity between the three noise groups.

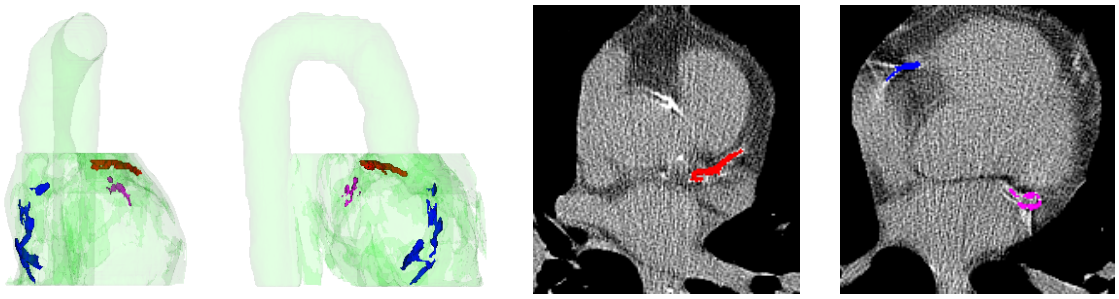
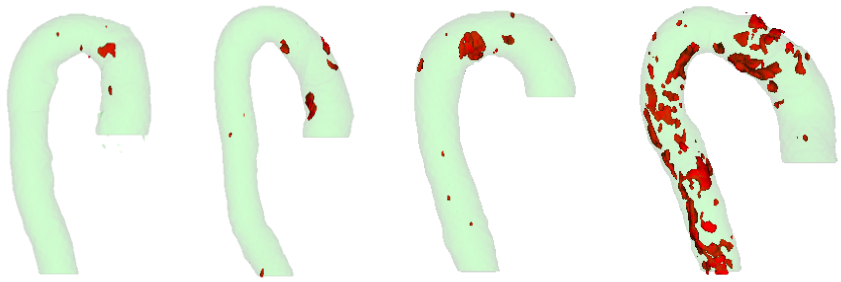

(a) (b) (c) (d)

Figure 2.31: Examples of correctly labeled CACs in a case. (a) and (b) show a coronal and a sagittal view with heart and aorta in light green; (c) and (d) show axial slices with labeled CACs. For colored CACs, LM+LAD CAC is in red, LCX CAC is in magenta, and RCA CAC is in blue.

For AC evaluation, the Pearson R is 0.99 for all three scores (AS, VS, MS) between automated and manual methods. The automated AS has a mean of 2937.0 (median=830.7, range=[0, 23160]). Figure 2.32 shows example cases with different levels of aortic calcification.

For PT and AA diameter evaluation, the Pearson R is 0.80 between reference pulmonary artery diameter and automated PT diameter, 0.93 between reference and automated AA diameter, and 0.81 between reference and automated diameter ratio. The Bland-Altman LOA results are listed in Table 2.16.



(a) (b) (c) (d)

Figure 2.32: Four example cases of AC (red) with aorta (light green) in sagittal view. The AS for the four cases are: (a) 420; (b) 1356; (c) 3038; (d) 18295.

Table 2.16. Bland-Altman LOA results of diameter comparison. PA=main pulmonary artery and Mean diff=Automated-Reference.

Automated	Reference	Mean diff	LOA	Range (Reference)
PT diameter	PA diameter	0.60	[-4.23, 5.43]	[19.16mm, 35.01mm]
AA diameter	AA diameter	-0.53	[-3.93, 2.85]	[22.90mm, 44.35mm]
PT/AA ratio	PA/AA ratio	0.03	[-0.12, 0.18]	[0.53, 1.10]

2.4.2 Qualitative evaluation results

The visual inspection results of cardiac segmentation for the large-scale qualitative evaluation are summarized in Table 2.17.

Table 2.17. Qualitative visual evaluation results for cardiac segmentation. For each grade category, both number of scans and percentage among all scans in this dataset are given.

Dataset	Number of Scans	Good	Acceptable	Unacceptable
VIA-ELCAP	46 (100%)	44 (96%)	2 (4%)	0 (0%)
LIDC	318 (100%)	293 (92%)	19 (6%)	6 (2%)
WTC	3766 (100%)	3247 (86%)	320 (8%)	199 (5%)
FY	2137 (100%)	1784 (83%)	200 (9%)	153 (7%)
QV	4672 (100%)	3854 (82%)	311 (7%)	507 (11%)
UN	9810 (100%)	8451 (86%)	520 (5%)	839 (9%)
All	20749 (100%)	17673 (85%)	1372 (7%)	1704 (8%)

2.5 Discussion

The aorta segmentation algorithm has good agreement with manually marked aorta boundaries (average DSC = 0.93, average boundary distance 1.39mm). Based on B_{pn}, the algorithm is biased towards over-segmentation. This is mainly because aorta is surrounded by

tissues with very similar intensity values in instances where there is only a small number of fat pixels adjacent to the aorta. Therefore, a small percentage of the image rays have not terminated correctly at the aorta surface. In very few cases (less than 0.5%), aorta segmentation fails in the centerline tracking stage due to 1) extremely high level of image noise (over 250HU inside soft tissue), or 2) abnormal anatomical structures (for example descending aorta on the right instead of the left side of the body).

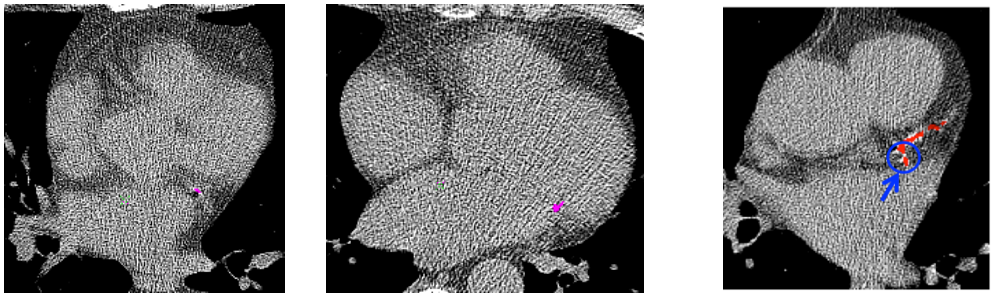
Segmentation of the PT proves to be more challenging than that of the aorta (average DSC = 0.88) because it is a much smaller region and is not as well-defined as the aorta. In addition, the segmentation of PT relies on the segmentation of aorta, therefore making it possible for aorta segmentation errors to propagate to PT. Similarly to aorta segmentation, the major cause of incorrect PT boundary is adjacent soft tissues within similar intensity values.

The heart region segmentation is not validated against manual markings because the segmentation target is the general heart region rather than the actual heart boundary, which is not visible in LDCT. Therefore as long as the segmented heart region covers all coronary arteries for the purpose of CAC evaluation, it is considered acceptable.

Detection of CAC is challenging in LDCT for a number of reasons including high noise level and heart motion artifacts. Techniques to reduce the impact of noise such as image filtering can also lead to the exclusion of small calcification regions. However, when comparing automated CAC with manual CAC, the automated CAC still has good agreement with reference CAC from low-dose scans (AS correlation 0.90, 33 cases correct risk categories and 7 cases one category off).

Automated CAC scores (all and by artery) are also compared to visual categorical scores from a much larger dataset with 1,359 LDCT scans. The Spearman R is 0.87 for all CACs. The

agreement for individual artery varies: it is the highest for LM+LAD calcification (Spearman R = 0.82) and the lowest for LCX calcification (Spearman R = 0.66). This is mostly because LCX contains the least amount of calcification and is more easily impacted by image filtering. The calcification in this artery also often locates at the branching point of the LM artery, making it difficult to separate the two regions. In addition, errors occur in a small number of cases where using geometric locations alone cannot distinguish CAC from other calcification in the proximity (see Figure 2.33 for correctly and incorrectly labeled LCX CAC).



(a) (b) (c)

Figure 2.33: Examples of CAC segmentation and labeling errors. (a) correctly labeled LCX CAC in magenta; (b) other calcification (in magenta) mistaken as LCX CAC; (c) LCX CAC (inside the blue circle) not separated from other LM+LAD CAC (in red).

Image noise level has an impact on both the calcification detection accuracy and sensitivity. Compared to images with lower noise, the high noise group (noise level above 110HU) has lower correlation and detection sensitivity overall. However, the difference is not obvious between low and medium noise groups and their difference is not statistically significant.

Both the AC and CVF segmentation have good agreement with manual markings. The Pearson R between automated and manual AC AS is 0.99. AC can be segmented using a relatively simple model because it is likely to be the only high-intensity structure on the surface of aorta. The DSC for CVF segmentation is 0.93. CVF under-segmentation occurs occasionally when the CVF anterior boundary has a complex shape and cannot be modeled accurately using a

rectangle.

It should also be noted that currently the calcification segmentation algorithms do not address the cardiac region metal implant issue or perform any implant detection. None of the scans in the evaluation datasets contains any major cardiac implants.

For the evaluation of PT and AA diameters, the automated and reference AA diameters have good agreement with good correlation (Pearson R = 0.93) and small LOA ([-3.93mm, 2.85mm]). The PT diameters and the diameter ratio have lower agreement (Pearson R = 0.80 and 0.81). This is mainly because the two PT measurements are taken at different locations and the PT diameter appears to be slightly larger than the main pulmonary artery diameter at the bifurcation level used clinically. It should also be noticed that there is likely to be a true difference in diameter measured from an axial image slice versus diameter measured from the 3D segmentation.

Through large-scale visual evaluation of cardiac segmentation, it is shown that the algorithms are robust when applied to a large dataset of heterogeneous CT scans. Among the 20,749 CT scans, 19,045 scans (92%) have good or acceptable result and for the 8% unacceptable scans, only about 3% is caused by incorrect cardiac segmentation itself. The remaining 5% out of the 8% belong to segmentation failures caused by incorrect segmentation of other organs that cardiac segmentation depends on, such as airway and ribs segmentation.

2.6 Conclusion

In conclusion, fully-automated algorithms have been developed for the segmentation of cardiac organs and the measurement of cardiac disease related image biomarkers in the context of low-dose chest CT scans. The cardiac organ segmentation includes the thoracic aorta, the pulmonary artery trunk, and the general heart region. The cardiac biomarker segmentation and

measurement includes the segmentation and measurement of CAC, AC and CVF. The aorta and PT diameters are also measured. The segmentation algorithms have shown both precision and robustness in the quantitative and qualitative evaluations. For quantitative evaluation, the average DSC is 0.93 for aorta, 0.88 for PT, and 0.93 for CVF. For qualitative evaluation on a large dataset of 20,749 chest CT scans, 17,673 scans (85%) have good cardiac segmentation results, 1,372 scans (7%) have acceptable cardiac segmentation results, and 1,704 scans (8%) have unacceptable results, the majority of which caused by unacceptable segmentation of other regions that cardiac segmentation depends on. The automatically measured biomarkers have shown good agreement with biomarkers extracted from manual annotations or visual ratings. The Pearson R between automated and manual AS is 0.90 for CAC and 0.99 for AC. The Spearman R between automated AS and visual scores is 0.87 for all CAC, and 0.82, 0.66, 0.72 for the individual artery groups. The Pearson R between automated and manual PT diameters and AA diameters is 0.80 and 0.93. The fully-automated cardiac region CAD algorithms provide the opportunity to evaluate cardiac disease risk level for a LCS population.

CHAPTER 3

LUNG DISEASE BIOMARKER MEASUREMENT

The primary purpose of LCS is the early-detection of pulmonary nodule, an indicator of lung cancer. LCS CT scans can also be used for the characterization of other lung region diseases such as the ILD and the thickening of pleural surface.

In contrast to CT scans obtained in a clinical setting for the evaluation of advanced lung diseases, participants of LCS are asymptomatic. Therefore it is very unlikely for advanced stage lung diseases to manifest in such scans. For this reason, the lung disease related algorithms in this dissertation are all designed for early-stage lung diseases. Three different disease measurements are presented:

1) Section 3.1 describes the characterization of small pulmonary nodule malignancy status and the impact of nodule size on the characterization outcome using LCS LDCT data;

2) Section 3.2 presents the identification of early-stage ILD by applying an automated algorithm to differentiate early-stage Usual Interstitial Pneumonia (UIP) from normal cases in LDCT;

3) Section 3.3 presents two algorithms to detect early-stage pleural surface thickening from LDCT. The first method measures the distance between pleural surface and ribs and the second method is based on the pleural surface shape.

3.1 Pulmonary nodule characterization

The major contribution of the nodule characterization algorithm presented in this dissertation is the construction of various training and evaluation datasets to assess the impact of nodule size on characterization outcome and to find the optimal training strategies (Reeves et al. 2016) [74]. The subsections are organized as follows:

1) section 3.1.1 presents the nodule dataset used in this study including the number of nodules

and their size distribution;

2) section 3.1.2 describes the image features computed from each nodule, which are used for machine learning classification;

3) section 3.1.3 describes the design of experiments and section 3.1.4 gives their corresponding results;

4) section 3.1.5 presents discussion and section 3.1.6 presents conclusion.

3.1.1 Construction of datasets

The whole dataset consists of nodule images from the two largest LCS studies, the International Early Lung Cancer Action Program (I-ELCAP) [6] and the National Lung Cancer Screening Trial (NLST) [75]. Malignant nodules are determined through pathological diagnosis and benign nodules are determined through pathological diagnosis or 2 years of no clinical change. Non-solid nodules are excluded from the dataset due to their very different visual presentation and slow growth rate. Only solid nodules or the solid component of part-solid nodules are included in the dataset. The whole dataset consists of a total of 736 nodules (412 malignant and 324 benign), among which 259 are from I-ELCAP and 477 are from NLST. Details of the whole dataset are listed in Table 3.1. All CT scans have a slice thickness of 3.2mm or less, with the majority of the scans less than or equal to 2.5mm (698/736=95%).

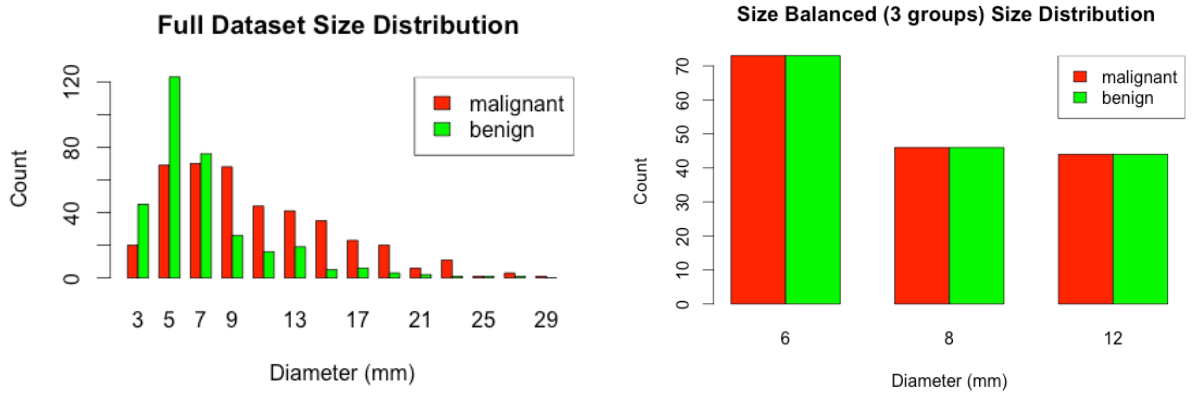
Table 3.1. Nodule numbers and equivalent diameters (in mm) statistics for the whole dataset and for I-ELCAP, NLST separately.

Dataset	Status	Number	Diameter Range (Median)
All	Malignant	412	[3.00, 29.14] (9.21)
	Benign	324	[3.11, 27.11] (5.84)
	Both	736	[3.00, 29.14] (7.33)
I-ELCAP	Malignant	167	[3.72, 29.14] (12.17)
	Benign	92	[3.25, 27.11] (9.47)
	Both	259	[3.25, 29.14] (11.28)
NLST	Malignant	245	[3.00, 27.60] (7.60)
	Benign	232	[3.11, 21.43] (5.15)
	Both	477	[3.00, 27.60] (6.15)

To assess the impact of nodule size distribution on characterization results, a size-balanced subset is constructed from the full dataset. This subset (GA) contains 163 malignant and 163 benign nodules. The two types of nodules are selected so that they have as similar size distribution (equivalent diameter) as possible. Details of GA are listed in Table 3.2. GA is further divided into 3 subsets of size-balanced malignant and benign nodules: G6 (diameter from 5 to 7 mm), G8 (diameter from 7 to 9 mm), and G12 (diameter from 9 to 14 mm). Details of these 3 subsets are also listed in Table 3.2. Figure 3.1 shows the size distribution for the whole dataset and GA.

Table 3.2. Nodule numbers and equivalent diameters (in mm) statistics for the GA set and for each binned subset of GA.

Dataset	Status	Number	Diameter Range (Median)
GA	Malignant	163	[5.01, 14.00] (7.21)
	Benign	163	[5.02, 13.91] (7.27)
	Both	326	[5.01, 14.00] (7.22)
G6	Malignant	73	[5.01, 6.98] (6.04)
	Benign	73	[5.02, 6.97] (6.06)
G8	Malignant	46	[7.04, 8.96] (7.65)
	Benign	46	[7.03, 8.84] (7.71)
G12	Malignant	44	[9.05, 14.00] (11.72)
	Benign	44	[9.02, 13.91] (11.41)



(a) (b)

Figure 3.1: Malignant and benign nodule size distribution. (a) whole dataset; (b) size-balanced subset GA divided into three subsets: G6, G8, G12.

3.1.2 Computation of image features

Each nodule in the dataset has its intensity CT image as well as a binary segmented nodule mask region using our previously published method (Reeves et al. 2006) [76]. Both the intensity and the mask images have been resampled to 0.25mm^3 isotropic resolution. Image features designed in our previous study (Jirapatnakul et al. 2007) [77] are computed based on the resampled intensity image and the nodule mask. There is a total of 46 image features, which can be grouped into 4 categories: morphological, density, surface curvature and margin gradient. A brief description is given for each category.

Morphological: they describe the 3D segmented nodule region shape and are derived from image moments (Prokop et al. 1992) [78]. The specific morphological features are: volume, surface area, volume to surface area ratio, compactness, sphericity, attachment ratio, length/width/height of the ellipsoid of inertia, ratios of the length/width/height to one another, the roll/pitch/yaw of the ellipsoid of inertia and the scale-normalized second-order morphological moment.

Density: they are derived from the gray-level intensity value of the image. The specific density features are: density mass, mean density, the standard deviation, skewness and kurtosis of the density histogram, length/width/height of the density-based ellipsoid of inertia, ratios of the

length/width/height of the density-based ellipsoid of inertia to one another and the scale-normalized second-order densitometric moment.

Surface curvature: they are computed from a smoothed polygonal tessellation of the segmented nodule region as described in (Jirapatnakul 2011) [79]. The following statistics of the curvature distribution are used as surface curvature features: mean, minimum, maximum, range, standard deviation, skewness and kurtosis.

Margin gradient: they measure the density changes occurring at the nodule margin, i.e. the segmented nodule boundary and the surrounding lung parenchyma. A surface tessellation of the nodule is computed and the surface normals are recorded. Image gradient samples are taken along each surface normal and the highest sample is recorded. The following statistics of the highest gradient samples are used as margin gradient features: mean, minimum, maximum, range, standard deviation, skewness and kurtosis.

3.1.3 Experiments and evaluation

For classification, five standard machine learning classifiers are used: the distance-weighted k-nearest-neighbors classifier (dwNN) (Dudani 1976) [80], the Support Vector Machine (SVM) classifier (Joachims 1999) [81] with a polynomial kernel (SVM-P), the SVM with a Radial Basis Function kernel (SVM-R), the logistic regression classifier (LOG) and the size threshold classifier (Size-C).

For dwNN, SVM-P, SVM-R and LOG classifiers, 5-fold cross validation is used for training and evaluation. In the training stage, 5-fold cross validation is used to divide the training set into training and validation sets for parameter optimization. The classifiers have probabilistic output and the final characterization outcome is represented by the average Receiver Operating Characteristic (ROC) curve (Fawcett 2006) [82] from the 5-fold as well as the area under this

ROC curve (AUC). For Size-C classifier, nodule size is the only feature and the ROC curve is computed by varying a size threshold through the size range in the dataset. At each step, a pair of sensitivity and (1-specificity) is obtained, based on which the ROC and AUC is computed. Two experiments are performed to validate the following hypotheses:

1) Nodule size has a major impact on characterization performance;

2) To characterize nodule status conditioned on a certain size range (i.e. evaluation on a size-balanced dataset), the classifier should be trained on a size-balanced dataset.

The two hypotheses are validated with the following two experiments listed in Experiment 3.1 and 3.2.

Experiment 3.1: Evaluation of nodule size impact on characterization outcome

1) Perform cross-validation on the whole dataset (unbalanced);
2) Perform cross-validation using size-unbalanced training dataset and size-balanced evaluation dataset GA;
3) Perform cross-validation using size-balanced dataset GA.

Experiment 3.2: Optimal training strategy when evaluating on a size-balanced dataset

1) Perform cross-validation using unbalanced training set and size-binned evaluation set (G6, G8, G12) (training and evaluation have the same size range);
2) Perform cross-validation using G6, G8, and G12;
3) Combining the best classifier trained on each individual bin and evaluated on GA through cross-validation;
4) Repeat steps 1-3 while replacing G8 and G12 with a combined larger bin G8+G12.

The two experiments are illustrated in Figure 3.2. For each experiment, besides the AUC value, each classifier is also compared to the AUC from the Size-C and the difference is defined as an Incremental AUC (IAUC) as it indicates the performance increase conditioned on a certain size. Both AUC and IAUC will be reported in the results. The p-values between the ROC curves (Size-C v.s. Others) are also computed to indicate the statistical significance of the IAUC

(DeLong et al. 1988) [83].

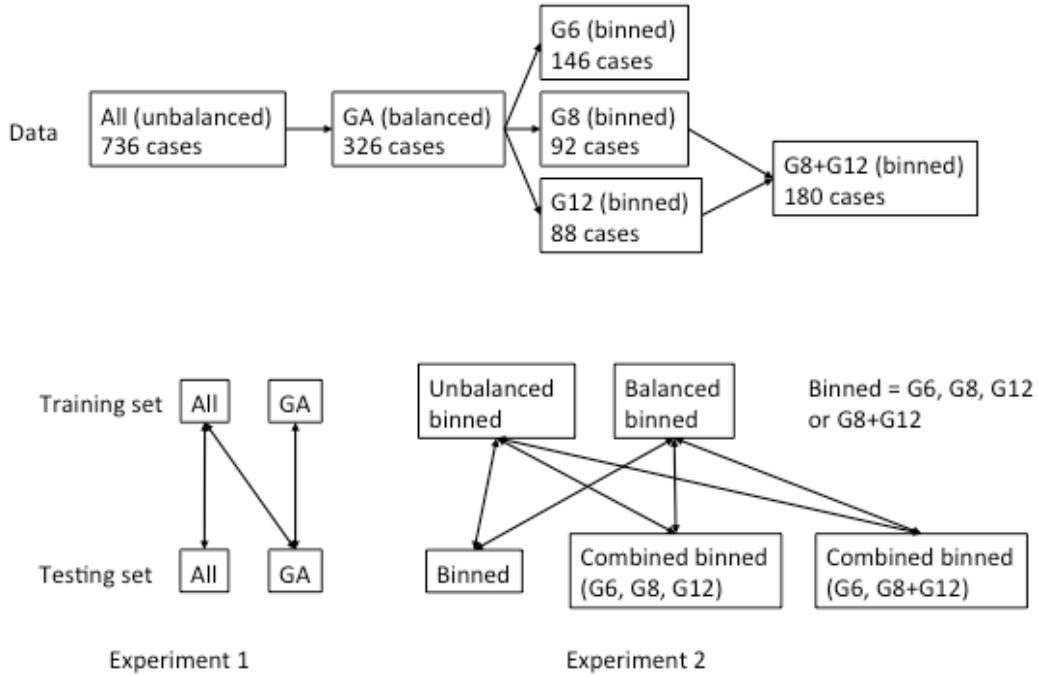


Figure 3.2: Organization of experiment 3.1 and 3.2.

3.1.4 Results

Results from Experiment 3.1 and 3.2 are summarized in Tables 3.3-3.5. The representative ROC plots are shown in Figure 3.3.

Table 3.3. Experiment 3.1 results. AUCs, IAUCs, and p-values under different training and evaluation conditions.

Training	Evaluation	Result	dwNN	SVM-P	SVM-R	LOG	Size-C
All (unbalanced)	All (unbalanced)	AUC	0.750	0.772	0.772	0.761	0.725
		IAUC	0.025	0.047	0.047	0.036	-
		p-value	=0.09	<0.001	<0.001	=0.15	-
All (unbalanced)	GA (balanced)	AUC	0.584	0.639	0.642	0.564	0.510
		IAUC	0.074	0.129	0.132	0.054	-
		p-value	=0.14	=0.01	=0.009	=0.11	-
GA (balanced)	GA (balanced)	AUC	0.700	0.708	0.699	0.624	0.510
		IAUC	0.190	0.198	0.189	0.115	-
		p-value	<0.001	<0.001	<0.001	=0.003	-

Table 3.4. Experiment 3.2 results. Best performance for the individual bins from GA under different training and evaluation conditions. All p-values are less than 0.001 and are not shown in the table.

Training	Evaluation	Results		Size-C AUC
G6 size range (unbalanced binned)	G6 (balanced binned)	AUC	0.646	0.546
		IAUC	0.100	
		classifier	dwNN	
G6 (balanced binned)	G6 (balanced binned)	AUC	0.691	0.546
		IAUC	0.145	
		classifier	LOG	
G8 size range (unbalanced binned)	G8 (balanced binned)	AUC	0.699	0.500
		IAUC	0.199	
		classifier	SVM-P	
G8 (balanced binned)	G8 (balanced binned)	AUC	0.759	0.500
		IAUC	0.259	
		classifier	SVM-P	
G12 size range (unbalanced binned)	G12 (balanced binned)	AUC	0.745	0.507
		IAUC	0.238	
		classifier	SVM-R	
G12 (balanced binned)	G12 (balanced binned)	AUC	0.759	0.507
		IAUC	0.252	
		classifier	SVM-P	
G8+G12 size range (unbalanced binned)	G8+G12 (balanced binned)	AUC	0.740	0.503
		IAUC	0.243	
		classifier	SVM-R	
G8+G12 (balanced binned)	G8+G12 (balanced binned)	AUC	0.780	0.503
		IAUC	0.277	
		classifier	SVM-P	

Table 3.5. Experiment 3.2 results. AUCs, IAUCs, and p-values after combining the best results from each individual bin. Two combination methods are used.

Training	Evaluation	Combining Method	Result		Size-C AUC
GA size range (unbalanced)	GA (balanced)	G6, G8, G12 (balanced 3 bins)	AUC	0.666	0.510
			IAUC	0.156	
			p-value	=0.002	
GA (balanced)	GA (balanced)	G6, G8, G12 (balanced 3 bins)	AUC	0.726	0.510
			IAUC	0.216	
			p-value	<0.001	
GA size range (unbalanced)	GA (balanced)	G6, G8+G12 (balanced 2 bins)	AUC	0.684	0.510
			IAUC	0.174	
			p-value	<0.001	
GA (balanced)	GA (balanced)	G6, G8+G12 (balanced 2 bins)	AUC	0.742	0.510
			IAUC	0.232	
			p-value	<0.001	

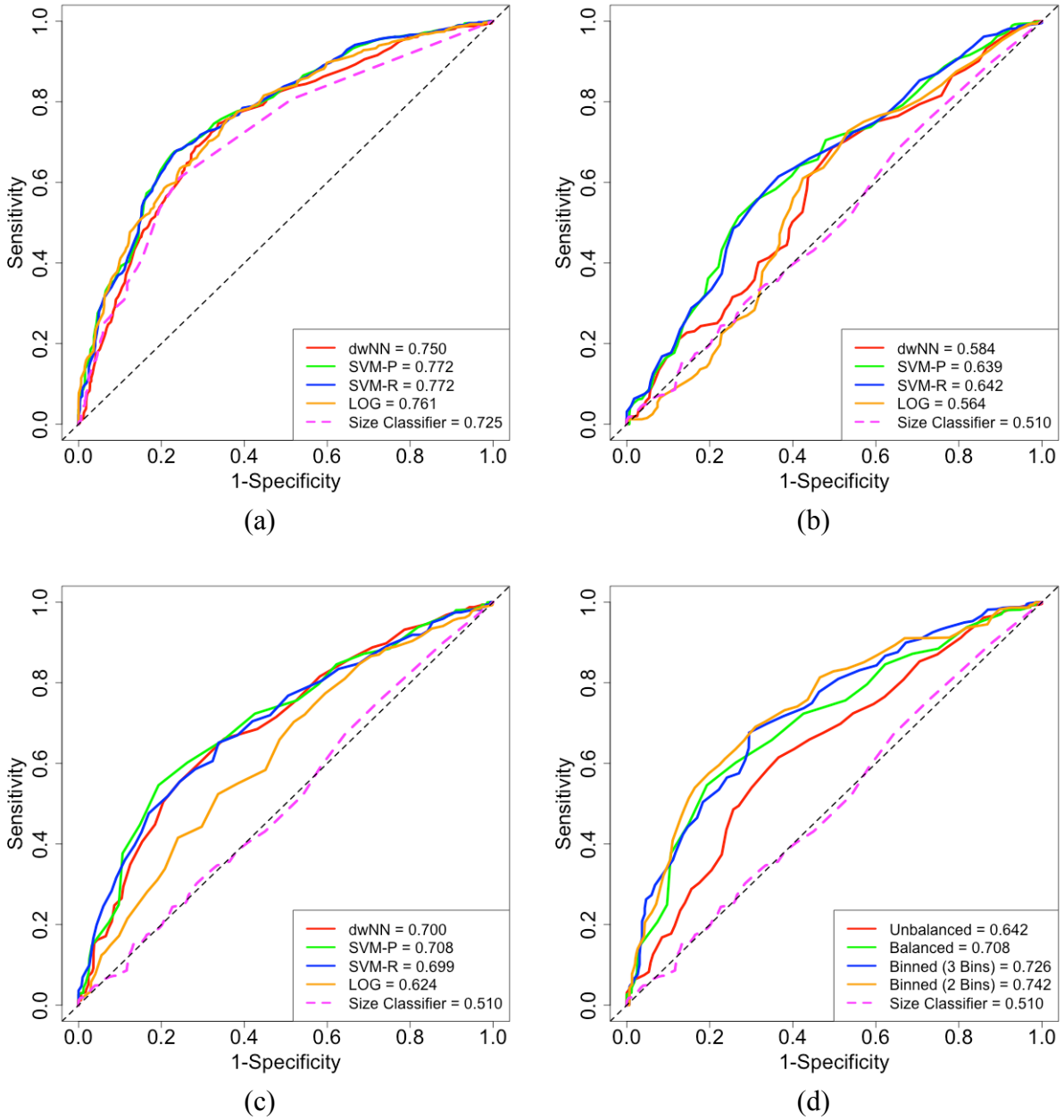


Figure 3.3: Result ROC curves. (a)-(c) are average ROC curves from experiment 3.1. (a) training and evaluation on all unbalanced data; (b) training on unbalanced and evaluation on balanced GA; (c) training and evaluation on balanced GA; (d) compares the best performance ROC curves from experiment 3.1 and 3.2 under different training and evaluation conditions.

3.1.5 Discussion

The two hypotheses have been validated through the experiments. First, nodule size plays an important role and without balancing for size between malignant and benign nodules, the

characterization outcome mostly reflects the natural size distribution. This can be seen in experiment 3.1 training and testing on all unbalanced data: the best classifier SVM-P and SVM-R (AUC=0.772) is just slightly better than using size alone (IAUC=0.047). Furthermore, this trained classifier has much lower AUC when evaluated on a size-balanced testing set (best AUC=0.642). However, the performance increases when the classifier is re-trained on a size-balanced set (best AUC=0.708), showing that balanced training is superior to unbalanced training when evaluated on a balanced testing set.

This is further proved in experiment 3.2. For evaluation on each individual bin, balanced training always outperforms unbalanced training even though both training methods use data from the same size range. The best overall performance when evaluated on GA is obtained by combining the best performance in individual bins (best AUC=0.742 after combining G6 with G8+G12).

It should also be noted that evaluation on a size-balanced set shows the performance of the features and classifiers conditioned on a fixed size range, as Size-C AUC for balanced evaluation set is close to random guessing (Size-C AUC close to 0.500). It is also observed that the larger size bins have better performance (G8 and G12 best AUC=0.759 while G6 best AUC=0.691). This could be caused by the fact that G6 nodules are much smaller in volume than G8 and G12 and have much higher percentage partial voxels, therefore making the image features less reliable.

There are some limitations to this study. First, there is no compensation for image noise, different imaging and scanner parameters; second, although a dataset with 736 nodules is considered fairly large for medical image related research, it is still far too small for many standard machine learning approaches.

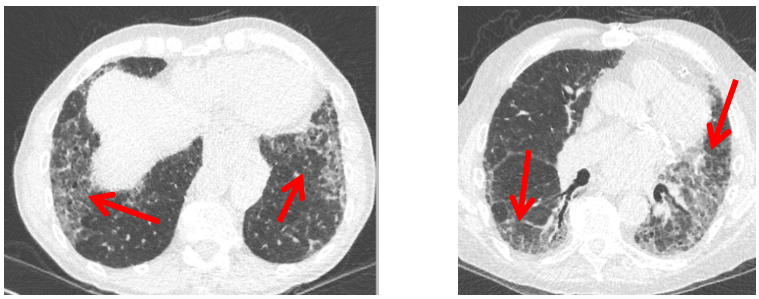
3.1.6 Conclusion

In conclusion, the size difference between malignant and benign nodules has a large impact on nodule status characterization and without taking this into consideration, characterization outcome may be overly optimistic. This can be avoided by evaluating on a size-balanced testing set.

Nodule characterization system should reflect the difference between malignant and benign nodules conditioned on a certain size range rather than the natural size difference between the two types. Performance of such system can be greatly improved by training on size-balanced nodules. Performance can be further improved by combining the best classifiers trained on size-balanced nodules within a small size range.

3.2 ILD identification

ILD is a group of lung parenchymal diseases. Common ILD types include usual interstitial pneumonia (UIP), nonspecific interstitial pneumonia (NSIP), hypersensitivity pneumonitis, and sarcoidosis. Some ILD types have known causes while many others are idiopathic. For example, hypersensitivity pneumonitis is often caused by exposure to allergens. Different ILD subtypes usually have different spatial distribution in terms of diseased lung parenchymal tissues (see Figure 3.4). Salvatore et al. (2016) [84] have shown that a common UIP pattern is subpleural, basilar-predominant lung fibrosis with traction bronchiectasis, ground-glass opacities, and reticulations without honeycombing and a possible NSIP pattern is central fibrosis.



(a) (b)

Figure 3.4: Example image slices from (a) an UIP and (b) an NSIP case showing different spatial distribution of diseased lung tissues (red arrows).

In this dissertation, a fully-automated algorithm (Xie et al. 2017) [85] has been developed to differentiate early-stage UIP from normal cases from LDCT scans. The algorithm captures the spatial distribution of lung fibrosis in the UIP cases by partitioning the lung into equal-volume subsections and the abnormality is characterized by a fractional high-density feature. In the subsections below, section 3.2.1 describes the lung partitioning algorithm and section 3.2.2 gives the fractional high-density feature. Section 3.2.3 presents the experiments and section 3.2.4 lists the corresponding results. Discussion and conclusion are given in section 3.2.5 and 3.2.6.

3.2.1 Equal-volume lung partitioning

The equal-volume lung partitioning method partitions the lung region into subsections with the same volume from three different directions: superior to inferior, anterior to posterior, and left to right. Then the partitions from the three directions are combined together to form the final partitions for each lung. This method is applied to the left and right lungs separately. The two lungs are segmented and separated using a previously developed algorithm (Kostis et al. 2003) [46]. The partitioning algorithm is described in Algorithm 3.1.

Algorithm 3.1 Lung partitioning algorithm

1) Perform equal-volume partitioning from superior to inferior direction using planes perpendicular to the image z-axis (e.g. intersection of coronal and sagittal planes) (see Figure 3.5 (a));
2) Perform equal-volume partitioning from anterior to posterior direction in each axial slice using planes perpendicular to the image y-axis (e.g. intersection of sagittal and axial planes). Partitions in each image slice are combined to form the anterior to posterior partitions for the entire lung (see Figure 3.5 (b)-(c));
3) Perform equal-volume partitioning from left to right in each axial slice by scanning lines of pixels parallel to the image x-axis (e.g. intersection of coronal and axial planes). Each line is partitioned into segments of the same length and partitions from all lines are combined together to form the left to right partitions for the entire lung (see Figure 3.5 (d)-(e));
4) Combine the partitions from step 1-3 resulting in the final subsection partitioning for the entire lung. If there is n partitions in each direction, the total number of subsections for each lung is n^3 (Figure 3.6 shows examples for n=3 and n=4).

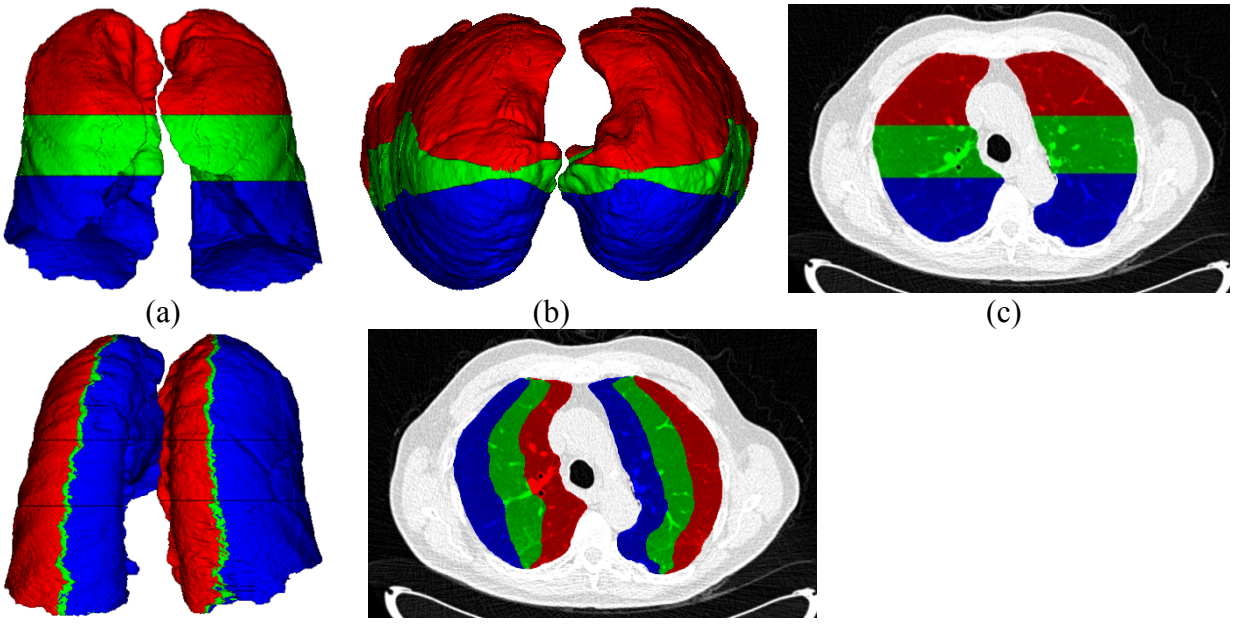


Figure 3.5: Lung partitioning visualization (3 partitions in each direction). (a) superior to inferior partitions in a coronal view; (b) shows the anterior to posterior partitions in a 3D axial view; (c) shows the anterior to posterior partitions in an axial slice; (d) shows the left to right partitions in a coronal view; (e) shows the left to right partitions in an axial slice.

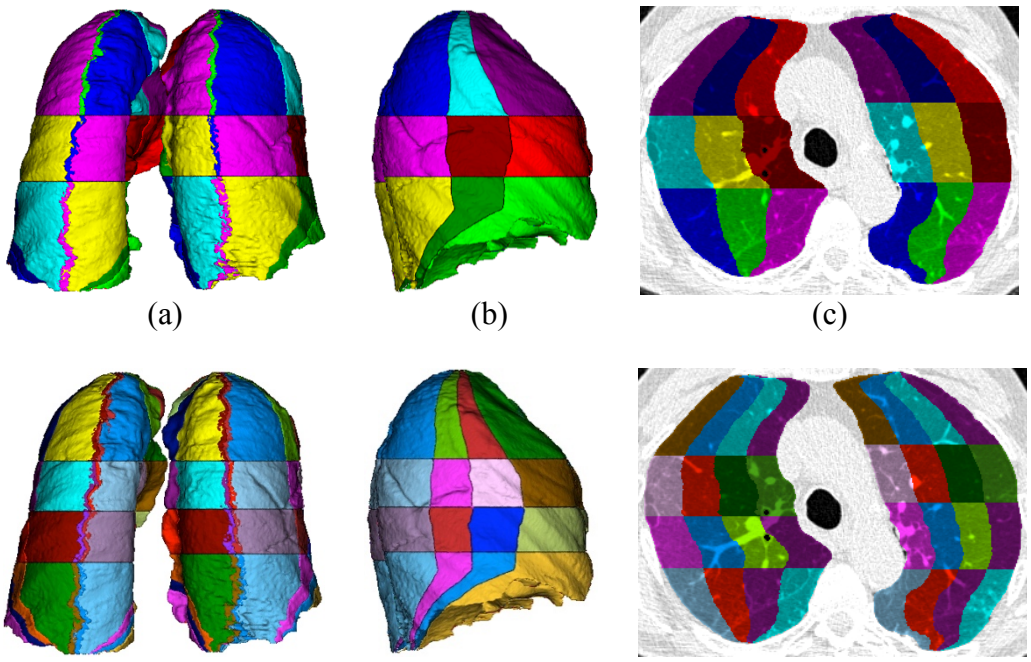


Figure 3.6: Combined partitions with 27 and 64 subsections. (a)-(c) show 27 subsections in a coronal view, a sagittal view, and an axial slice; (d)-(f) show 64 subsections in a coronal view, a sagittal view, and an axial slice.

Image features can be computed from each subsection and then combined together. The equal-volume strategy ensures that each subsection contains the maximum number of pixels for feature computation. Another lung partitioning based ILD characterization method has been developed by Depeursinge et al. (2015) [86]. Compared to their method, this algorithm is more general and allows for different numbers of equal-volume partitions.

3.2.2 Computation of fractional high-density features

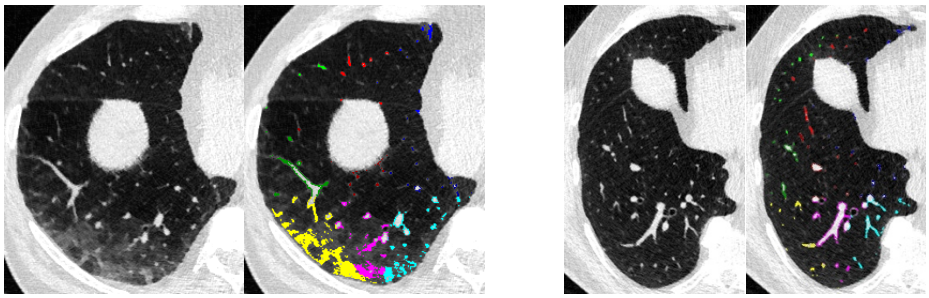
For early-stage UIP cases, the abnormal lung tissues do not typically exhibit very distinguishable texture features. Therefore, a classic fractional high-density feature is used to capture the UIP fibrosis regions. The high-density range of [-600HU, -250HU] is defined by Lederer et al. (2009) [87]. This range characterizes subclinical parenchymal lung disease while excluding more dense areas such as complete atelectasis, medium to large blood vessels and pulmonary nodules. In each subsection i, the fractional high-density measurement h_i is computed

as the fraction of pixels in that subsection within the range of [-600HU, -250HU]. Algorithm 3.2 describes the combination of h_i from all subsections for each case.

Algorithm 3.2 ILD feature vector computation for the whole lungs

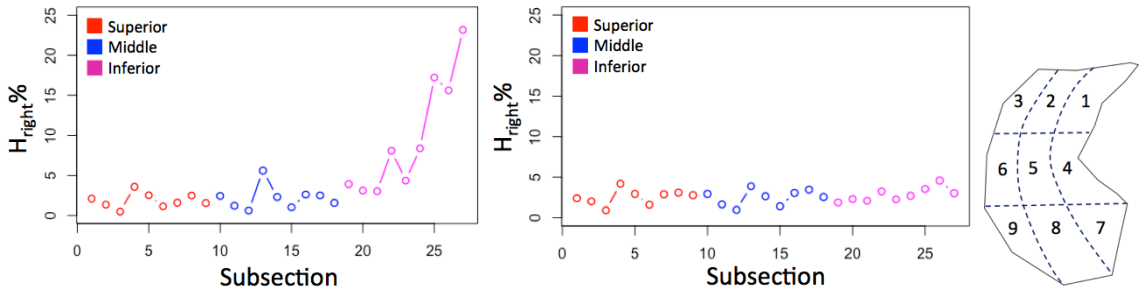
1) Denote the number of subsections in each lung as n. Denote the left and right lung feature vectors as H_{left} and H_{right}:
   $H_{left} = \{h_{left1}, h_{left2}, \dots, h_{leftn}\}$ (3.1)
   $H_{right} = \{h_{right1}, h_{right2}, \dots, h_{rightn}\}$ (3.2)
2) Combining H_{left} and H_{right} through concatenation, resulting in a feature vector H_{all} of length 2n:
   $H_{all} = \{H_{left}, H_{right}\}$ (3.3)
3) Combining H_{left} and H_{right} through selecting the maximum value between the two lungs, resulting in a feature vector H_{max} of length n:
   $H_{max} = \{\max(h_{left1}, h_{right1}), \max(h_{left2}, h_{right2}), \dots, \max(h_{leftn}, h_{rightn})\}$ (2.4)

Example h_i measurements in a UIP lung and a normal lung are shown in Figure 3.7. It can be observed that the high-density pixel distribution is different for the UIP and the normal case. This is also demonstrated in Figure 3.8 when plotting the h_i values in the 27 subsections for the UIP and the normal lung.



UIP UIP h_i pixels Normal Normal h_i pixels

Figure 3.7: High-density (h_i) pixels in an UIP lung and a normal lung. Each color corresponds to a different subsection (27 subsections in total).

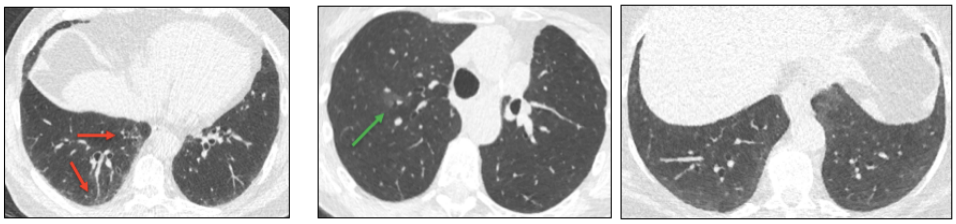


(a) UIP H_{right} (b) Normal H_{right} (c) Ordering

Figure 3.8: The distribution of H_{right} for the same UIP and normal case shown in Figure 3.7 (27 subsections in each lung). (a) UIP lung H_{right} distribution; (b) normal lung H_{right} distribution; (c) subsection number ordering in each layer (superior, middle and inferior).

3.2.3 Experiments and evaluation

The evaluation dataset consists of 56 normal cases and 51 UIP cases from different subjects. All CT scans are low-dose non-contrast and thin-slice (0.5-1.25mm slice thickness). All scans have been visually inspected by a radiologist to establish their status as normal or UIP. Normal cases do not exhibit ILD related patterns. However, they may contain other types of abnormalities such as nodules or lung atelectasis (see Figure 3.9).



(a) (b) (c)

Figure 3.9: Example UIP and normal cases in the evaluation dataset. (a) UIP case (red arrows indicate abnormal regions); (b) normal case with a ground glass opacity indicated by the green arrow; (c) normal case with lung atelectasis.

The automated lung partitioning and fractional high-density feature extraction method is applied to each case to compute the H_{all} and H_{max} features. H_{all} and H_{max} are evaluated separately. Lung partitioning number n is also varied: n = 1, 8, 27, and 64. An SVM classifier with a radial basis function (RBF) kernel is used to classify the cases into normal or UIP using these features and ten-fold cross-validation is used. The average AUC is used as the evaluation metric. The statistical significance of the difference between the two ROC curves is evaluated using the

DeLong test [83].

3.2.4 Results

The H_{left} and H_{right} distributions for the UIP cohort and the normal cohort without partitioning (partition number n = 1) are shown in the histograms in Figure 3.10. After partitioning each lung into 27 subsections, the h_{lefti} and h_{righti} averaged over the entire cohort are shown in Figure 3.11. Tables 3.6-3.7 show the AUC values for each partition number n and for both H_{all} and H_{max}. The p-values between no partition (n = 1) and with partitions (n > 1) are also computed and listed in Tables 3.6-3.7. For the same n number, ROC using H_{all} is compared to ROC using H_{max} and there is no statistically significant difference between them (p-value > 0.05).

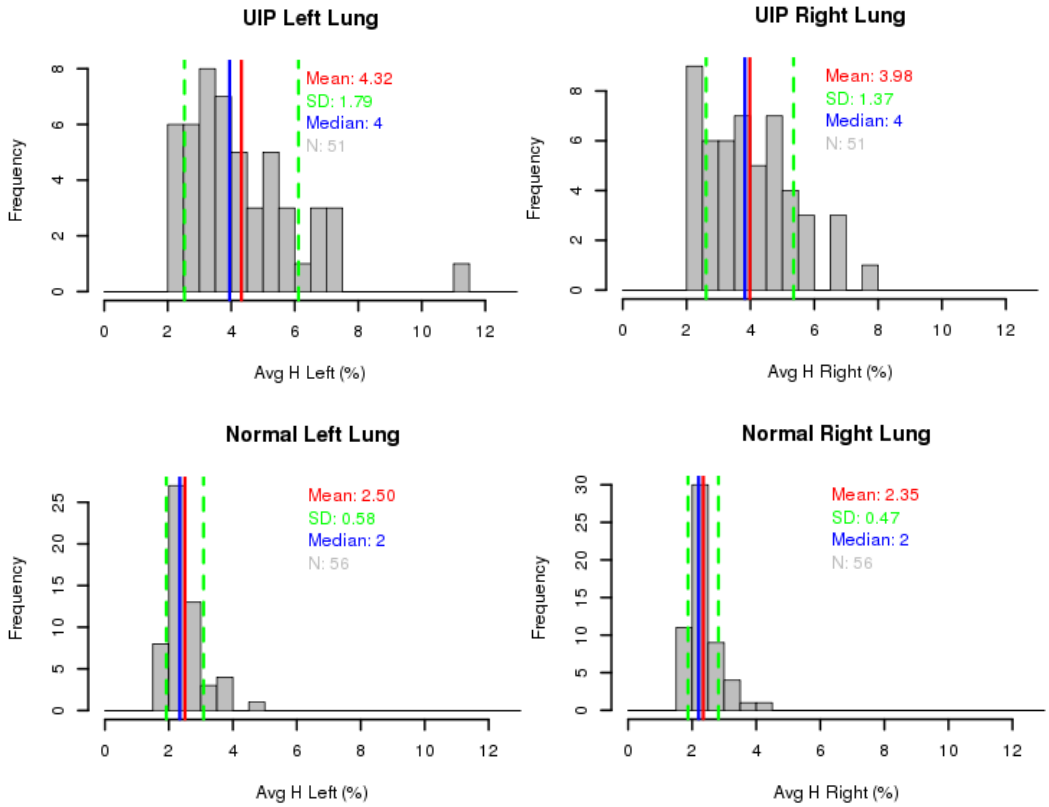


Figure 3.10: Histogram of H_{left} and H_{right} without partitioning for the UIP and the normal cohort.

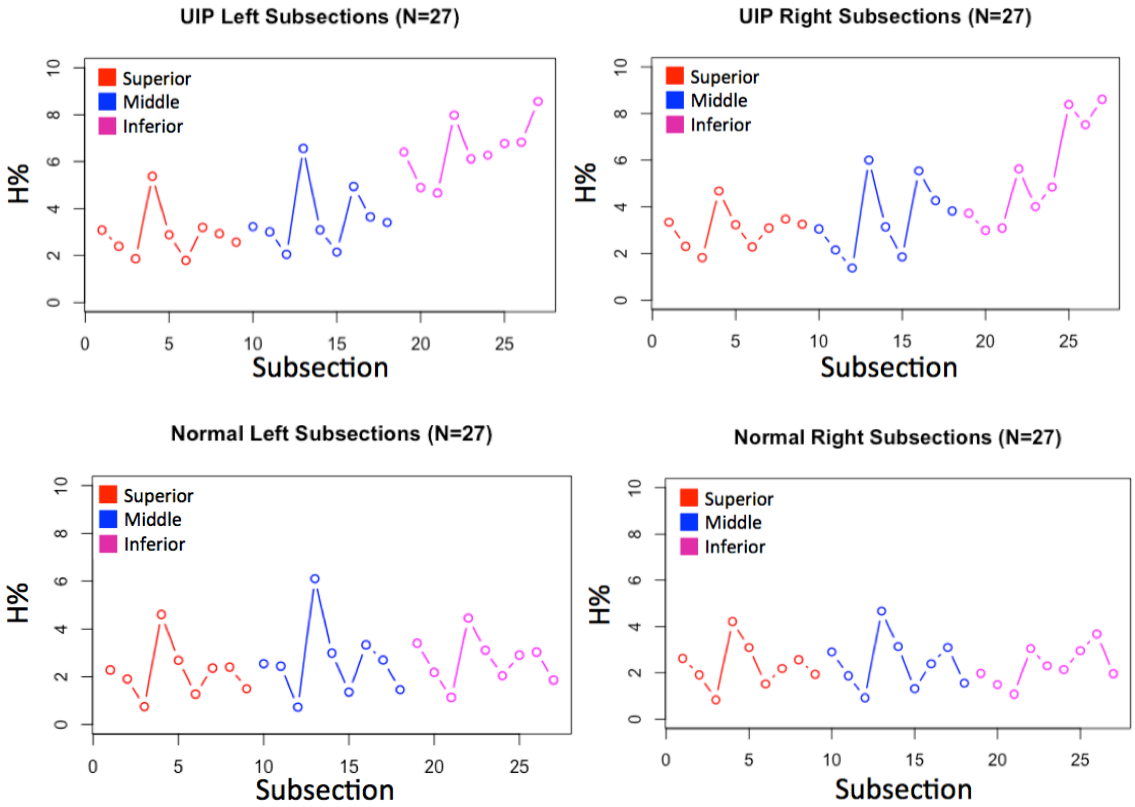


Figure 3.11: Averaged h_{lefti} and h_{righti} for the UIP and the normal cohort (27 subsections in each lung).

Table 3.6. Classification average AUC values using H_{all} with different numbers of partitions n in each lung.

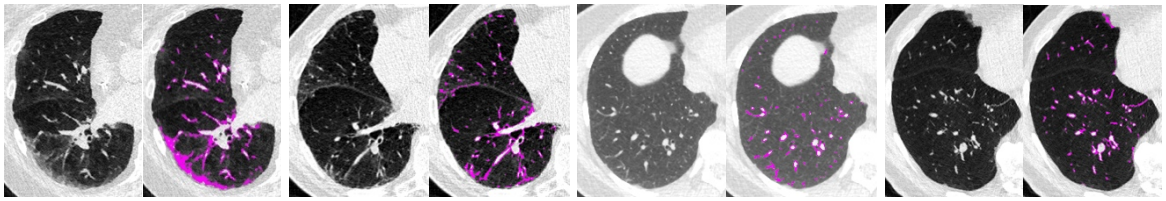
Partition number n (Feature length = 2n)	AUC	p-value (compared to n=1)
n= 1	0.85	-
n = 8	0.92	0.14
n = 27	0.95	0.026
n = 64	0.94	0.035

Table 3.7. Classification average AUC values using H_{max} with different numbers of partitions n in each lung.

Partition number n (Feature length = n)	AUC	p-value (compared to n=1)
n = 1	0.85	-
n = 8	0.92	0.12
n = 27	0.93	0.056
n = 64	0.93	0.064

3.2.5 Discussion

The automated algorithm is able to robustly separate early-stage UIP from normal cases with a highest AUC of 0.95. The fractional high-density measurement is useful in capturing the abnormal tissues in UIP cases without capturing major vessels or nodules. Although the transition region between vessels and lung parenchyma is also included by the high-density range, this happens evenly for both UIP and normal cases and do not seem to have a major impact on the outcome (see Figure 3.12).



UIP case 1 UIP case 2 Normal case 1 Normal case 2

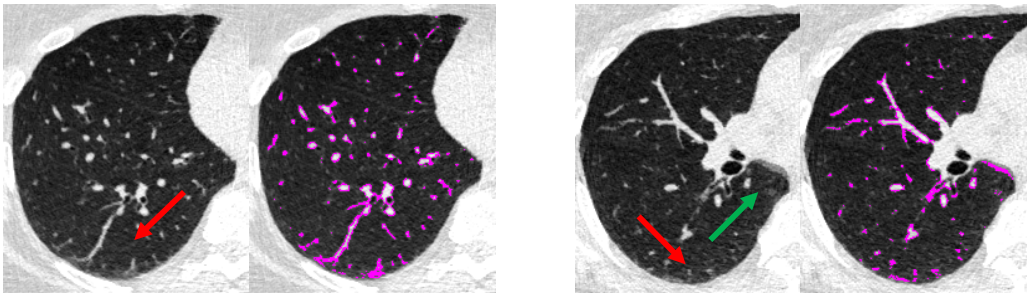
Figure 3.12: Example UIP cases and normal cases with high-density pixels in magenta.

The spatial distribution of lung fibrosis is different for the UIP and the normal cohort as shown in Figure 3.11. While normal cases have very similar h_i values across all subsections, the UIP cases have a much higher h_i value in the inferior and posterior regions. This phenomenon can be captured through lung partitioning. When partition number increases from n = 1 to n = 27, the AUC increases from 0.85 to 0.95 with statistical significance (p-value = 0.026). However, when n is further increased to 64, the AUC values does not further increase. One potential reason is that when n further increases, the number of pixels in each subsection decreases, making the measurement less robust.

There is not a statistically significant difference between using H_{all} and using H_{max} although H_{all} leads to higher AUC values. In fact, based on the h_i distribution shown in Figure 3.11, the UIP patterns appear to be somewhat bilateral.

In a small number of regions, the h_i measurement has incorrectly captured certain regions in normal cases such as the lung atelectasis regions (see Figure 3.13 (a)) and the dense region

caused by heart motion (see Figure 3.13 (b)).



(a) (b)

Figure 3.13: Example normal cases with high h_i values (high-density pixels in magenta). (a) lung atelectasis (red arrow); (b) ung atelectasis (red arrow) and heart motion (green arrow).

This algorithm has the potential to be extended to distinguish between different ILD subtypes such as UIP versus NSIP in the future when more data become available. Using the established lung partitioning model, additional image features can be easily incorporated.

3.2.6 Conclusion

In conclusion, a fully-automated algorithm has been developed to distinguish early-stage UIP cases from normal cases in LDCT. The algorithm partitions the lung region into equal-volume subsections and computes a fractional high-density feature for each subsection. It is able to distinguish early-stage UIP from normal cases with a highest AUC of 0.95 (with 27 subsections). Using lung partitioning over not partitioning shows a statistically significant improvement in AUC value. This algorithm may be used to aid the diagnosis of early-stage UIP in the context of LDCT.

3.3 Pleural thickness measurement

Pleural thickening refers to the thickening of visceral pleura and can be induced by exposure to asbestos (Gevenois et al. 1998) [37]. It is often found in patients with malignant pleural mesothelioma, a usually fatal disease. However, the target in this study is to detect the early-stage subtle pleural thickening regions rather than the much more severe pleural thickening manifesting in malignant mesothelioma cases.

In this study, two different methods are designed to capture potential pleural thickening in a cohort exposed to the Libby vermiculite. The subsections are arranged as follows:

1) section 3.3.1 presents the first pleural thickening measurement method that measures the sampled distances between ribs and pleural surface;

2) section 3.3.2 presents the second pleural thickening measurement method that measures the volume of concave regions on the pleural surface;

3) section 3.3.3 describes experiments and evaluation method and section 3.3.4 gives the corresponding results;

4) section 3.3.5 and 3.3.6 presents discussion and conclusion.

3.3.1 Measurement using distance between pleural surface and ribs

Distance between ribs and pleural surface is used as a surrogate measurement to the actual pleural thickness for two reasons: first, in early-stage pleural thickening cases, the abnormal region is very small and subtle and it may be difficult to measure them directly; second, it is observed that early-stage pleural thickening cases usually have an overall increased distance between pleural surface and ribs due to pleural abnormality (see Figure 3.14). Therefore, instead of trying to directly capture small abnormal regions, the overall distances are measured. The algorithm first obtains the evenly sampled minimum distances between pleural surface and ribs. Then samples in unreliable regions are eliminated where there is likely to be a large space between ribs and pleural surface due to natural anatomical structures. The algorithm is listed in Algorithm 3.3.

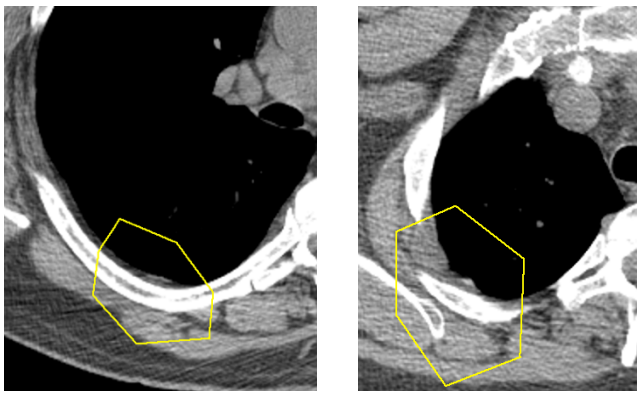
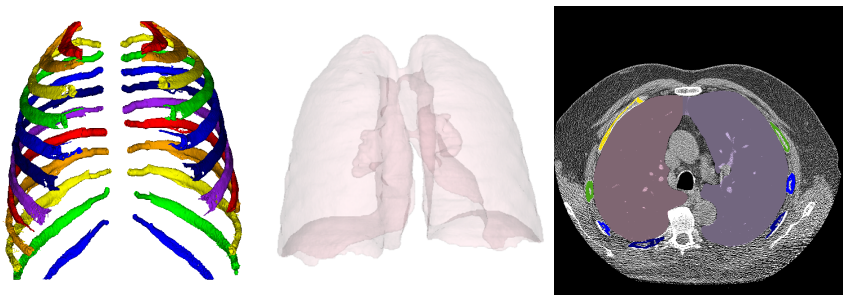


Figure 3.14: Pleural thickening in two cases outlined by a radiologist (yellow), both of which manifest as an enlarged distance between the rib and the pleural surface.

Algorithm 3.3 Distance measurement between ribs and lungs

1) Obtain segmented ribs and lungs using a previously published method (Lee et al. 2010) [48] (see Figure 3.15) and cast image rays axially from lung center towards pleural surface;
2) Retain rays that intersect with a rib and denote the distance between the intersection point and pleural surface as R_{ijk}, where i is the CT slice number, j is the ray angle ([0°, 360°] with a step size of 4°), and k is the rib label (see Figure 3.16);
3) For each k at each j, select the ray R_{ajk} that minimizes the distance as shown in Figure 3.17:
$$R_{ajk} = \min\{R_{ijk}\}, i \in [1, n] \quad (2.5)$$
4) Exclude measurements in the most superior section of each lung, adjacent to vertebrae, or in the most inferior left lung close to the diaphragm (see Figure 3.18 for exclusion examples);
5) Obtain the remaining distance measurements as shown in Figure 3.19 and compute their statistics: mean D_{mean} and standard deviation D_{sdev}.



(a) (b) (c)

Figure 3.15: Visualization of segmented and labeled ribs and lungs. (a) coronal view of labeled ribs in different colors, (b) coronal view of the segmented lungs; (c) both ribs and lungs in an axial slice overlaid on an intensity image in different colors.

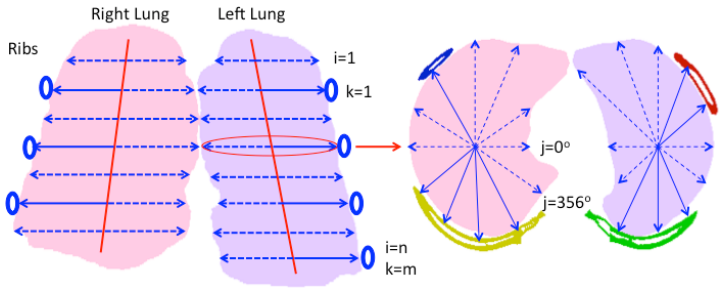


Figure 3.16: Ray casting model. Image rays intersecting with a rib are shown in solid blue lines while other rays are in dashed lines. Lungs are in different shades of pink and ribs are colored individually. Image slice (i), rib label (k) and angle j are also shown.

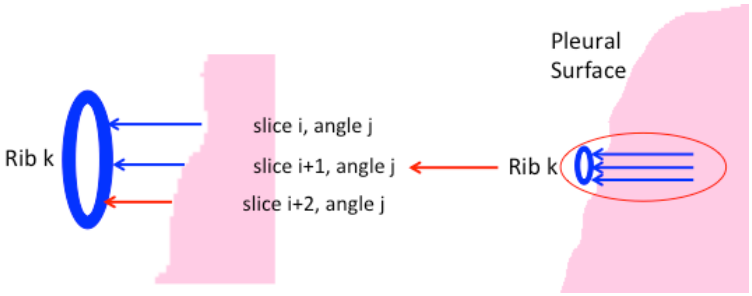
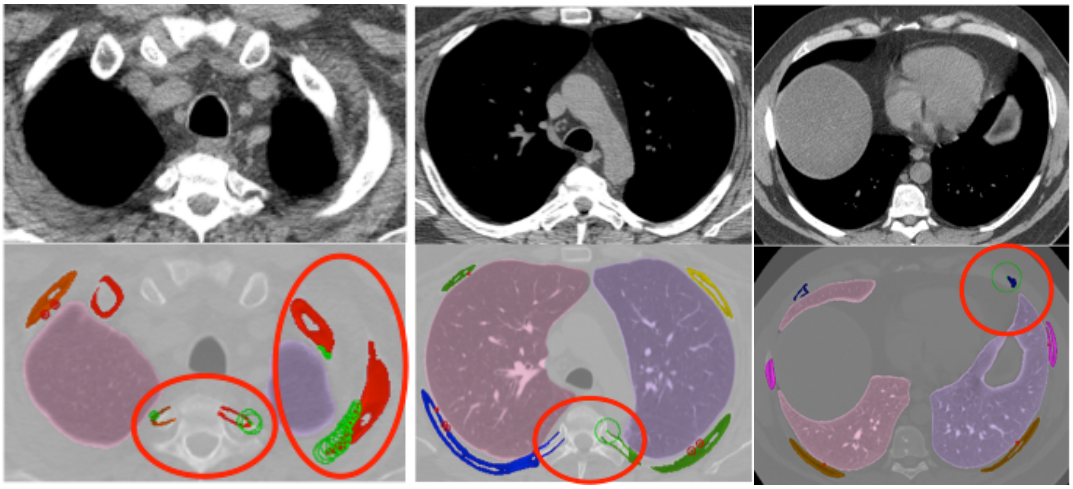
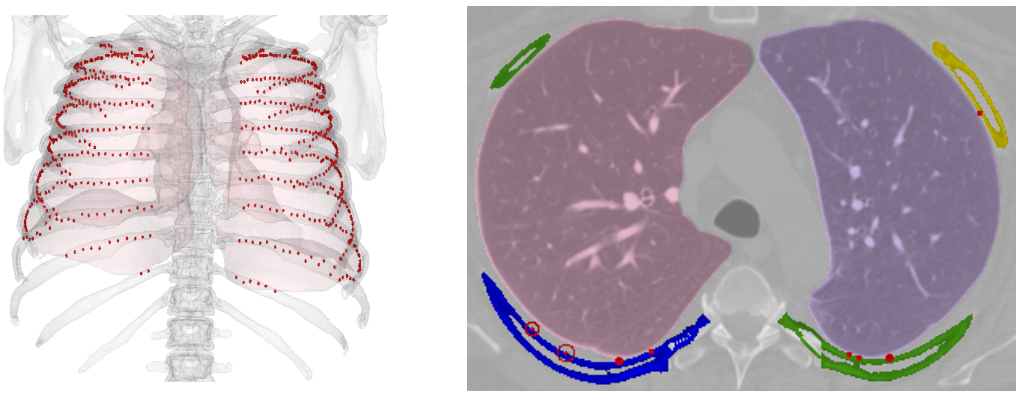


Figure 3.17: Demonstration of ray selection. Among the three rays intersecting with rib k at angle j, only the ray with the minimum distance is selected.



(a) (b) (c)

Figure 3.18: Examples of excluded distance measurements. Red circles indicate the general regions and green circles indicate the excluded measurements (radius of the circle is the actual distance). (a) exclusion in the superior lung and vertebrae adjacent regions; (b) exclusion in the vertebrae adjacent regions; (c) exclusion in the inferior left lung close to the diaphragm.



(a) (b)

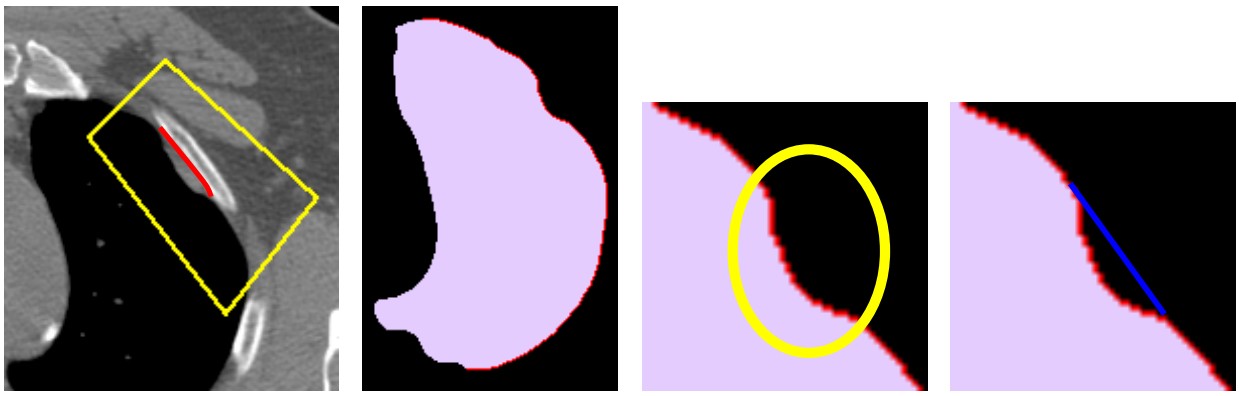
Figure 3.19: Visualization of final distance measurements in red circles. (a) and (b) show their locations relative to the ribs and lungs in coronal and axial views.

3.3.2 Measurement using concave regions on the pleural surface

While the first method in section 3.3.1 focuses on the overall increase of distances between pleural surface and ribs, this method measures pleural thickening by detecting and extracting concave regions on the pleural surface. This method is designed to capture the more visible pleural thickening that causes a measureable concavity on the pleural surface (see Figure 3.20 (a)). The algorithm is listed in Algorithm 3.4. It first detects and measures the concave region volume using a surface tracing and convex hull method. Then within all the concave regions, a subset of regions not adjacent to ribs is extracted and their volume also measured. This is to avoid the natural pleural surface indentation caused by the ribs. Both volume measurements are used in evaluation.

Algorithm 3.4 Concavity measurement on the pleural surface

1) Trace the costal pleural surface on the segmented lung from anterior to posterior (see Figure 3.20 (b)-(c));
2) Compute a convex hull on the traced costal surface;
3) The differences before and after the convex hull operation are considered to be the concave regions as shown in Figure 3.20 (d);
4) Extract concave region pixels within the soft tissue intensity range and compute their volume as Vol_A;
5) Exclude concave regions less than 5mm from the segmented ribs and re-compute their volume Vol_S.



(a) (b) (c) (d)

Figure 3.20: Extraction of a concave region on the pleural surface. (a) target concave region; (b) traced pleural surface in a red line; (c) a zoomed-in concave region in (b) (yellow outlined); (d) detected concavity (region bounded by blue and red lines).

3.3.3 Experiments and evaluation

The evaluation dataset consists of 100 chest CT scans of subjects exposed to Libby vermiculite. The scans are from different subjects, non-contrast and have a slice thickness of 1.25mm or less. The lung region in each scan is partitioned into 6 subsections: Left Upper lung (LU), Left Middle lung (LM), Left Lower lung (LL), Right Upper lung (RU), Right Middle lung (RM) and Right Lower lung (RL), where LU, LM and LL are the superior, middle and lower lung regions with equal volume and the same applies to RU, RM and RL. The two pleural thickness measurements are performed in the 6 subsections as well as in each lung. Each subsection has a visual rating by a radiologist. The rating ranges from 1 to 4 and the higher the rating is, the more severe pleural thickening in that subsection is. A rating of 1 indicates no or minimal abnormality.

Two experiments are performed: the first experiment evaluates the agreement between the computer measurements with the visual ratings by each subsection and the second experiment evaluates the agreement by each lung, in which the median rating of the 3 subsections of each lung is chosen to be the rating of that lung. Both Pearson correlation coefficient R and Spearman correlation coefficient R are computed in the two experiments.

3.3.4 Results

Results from the two experiments are summarized in Tables 3.8-3.9. The visual inspection statistics is summarized in Table 3.10. Method 1 is the pleural surface and ribs distance measurement and Method 2 is the concavity volume measurement. Experiment 1 evaluates the agreement by subsection and Experiment 2 evaluates the agreement by each lung.

Table 3.8. Pearson R from pleural thickening experiments.

Pearson R	Method 1		Method 2	
	D_{mean}	D_{sdev}	Vol_A	Vol_S
Experiment 1	0.32	0.35	0.60	0.60
Experiment 2	0.33	0.40	0.69	0.74

Table 3.9. Spearman R from pleural thickening experiments.

Spearman R	Method 1		Method 2	
	D_{mean}	D_{sdev}	Vol_A	Vol_S
Experiment 1	0.21	0.23	0.33	0.35
Experiment 2	0.21	0.24	0.37	0.36

Table 3.10. Visual inspection statistics (mean, median and standard deviation) for each subsection, each lung and the two lungs.

Location	Mean	Median	Standard deviation
LU	1.29	1	0.66
LM	1.31	1	0.71
LL	1.23	1	0.63
RU	1.39	1	0.85
RM	1.36	1	0.79
RL	1.24	1	0.62
Left lung	1.28	1	0.66
Right lung	1.33	1	0.76
Lungs	1.30	1	0.71

3.3.5 Discussion

In both experiment 1 and 2, method 2 (volume of concave regions) has higher agreement with visual ratings than method 1. However, neither method has shown a strong correlation with visual ratings. It should be noted that method 2 targets regions more visible to human observers while it is difficult for observers to identify the subtle changes targeted by method 1. Therefore

method 2 is likely to have a higher agreement with human observers. Another potential reason for the overall low agreement is that the majority of the cases have minimal visually detectable pleural thickening as shown in Table 3.10. Therefore without the information of patients' medical history or clinical diagnostic outcome, it is difficult to judge whether the computer algorithms are measuring anything useful for diagnosis in these early-stage pleural thickening cases.

Overall, both methods are able to correctly capture many abnormal regions (see Figure 3.21 upper row). However, a common false positive for both methods is the large distance or indentation caused by ribs (see Figure 3.21 lower row). These regions have no visual difference compared to the true pleural thickening regions and cannot be reliably excluded by the current methods.

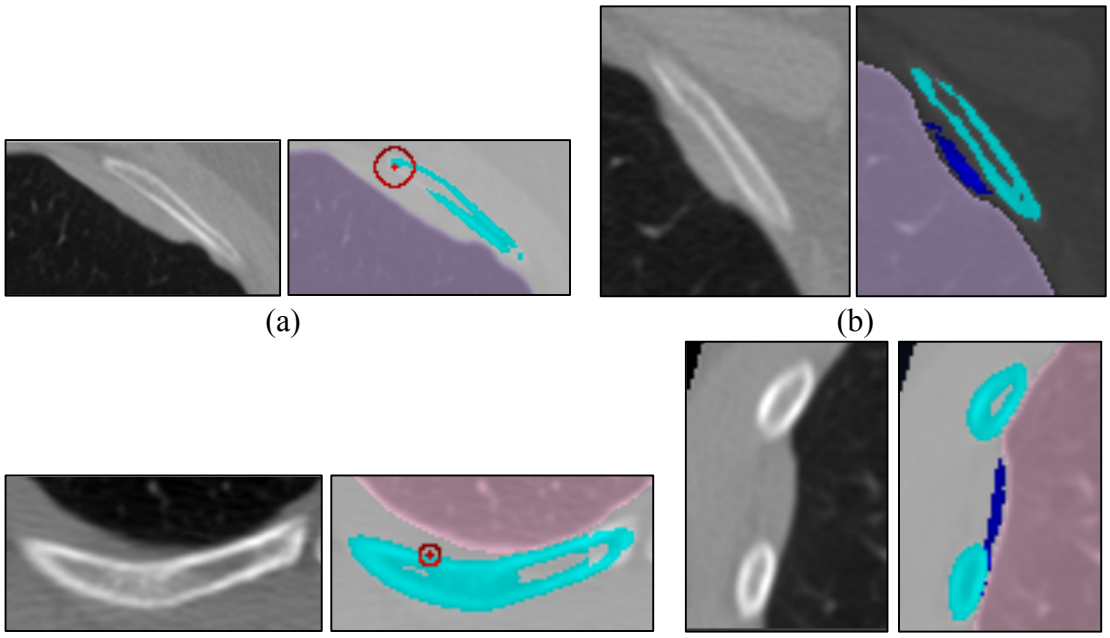
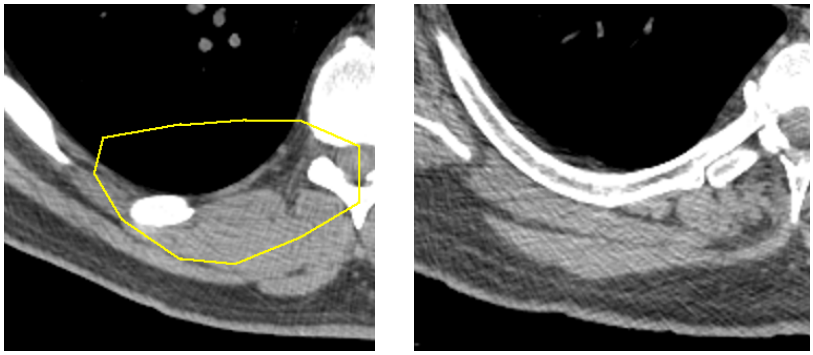


Figure 3.21: Correctly identified pleural thickening regions (upper row) and normal regions incorrectly identified as pleural thickening (lower row). (a) and (c) are identified by method 1 (red circles); (b) and (d) are identified by method 2 (blue regions).

In addition, neither method is designed to capture the pleural thickening adjacent to the vertebrae since both methods explicitly exclude measurements in this region. This exclusion is

performed because the pleural surface adjacent to the vertebrae has a very similar appearance to the real pleural thickening regions such as a large indentation and a large space between pleural surface and ribs (see Figure 3.22). More advanced algorithm is needed to measure pleural thickening in this region.



(a) (b)

Figure 3.22: Pleural surface regions adjacent to vertebrae in two cases with similar visual presentations. (a) pleural thickening; (b) normal (through visual rating).

3.3.6 Conclusion

In conclusion, two fully automated computer methods have been designed to detect early-stage pleural thickening by measuring the pleural surface to rib distances and the volume of pleural surface concave regions. Evaluated on a cohort of subjects exposed to Libby vermiculite, both methods show some correlation with visual ratings with the concavity measurement having higher correlation (Pearson R=0.75 and Spearman R=0.36). To evaluate the potentials of the algorithms, a much larger dataset with documented clinical diagnostic outcomes is needed.

CHAPTER 4

IMAGE QUALITY MEASUREMENT

The image quality (noise and calibration) of a CT scan has a direct impact on the performance of CAD algorithms. Noise impacts the detectability of image biomarkers. For instance, in Chapter 2 it has been shown that the detectability of CAC significantly decreases when the heart region noise level increases from below to over 110HU. Calibration impacts thresholding based algorithms and biomarker measurements. For example, one of the most common emphysema measurements, emphysema index, is defined to be the fractional pixels in the lung region below the intensity threshold of -950HU. Therefore, this index will be inaccurate if the CT scan is not properly calibrated. Image quality can also impact the visual reading precisions of a CT scan by human observers.

In this dissertation, a fully-automated method to evaluate CT image quality is presented. It takes advantage of the pre-segmented regions such as the aorta and is a natural extension of the previously developed CAD algorithms in Chapters 2-3. In the following sections, section 4.1 describes the automated extraction of regions used for quality measurement; section 4.2 describes the evaluation methods and the evaluation datasets for the automated quality measurement; section 4.3 gives the evaluation results; section 4.4 and section 4.5 gives the discussion and conclusion.

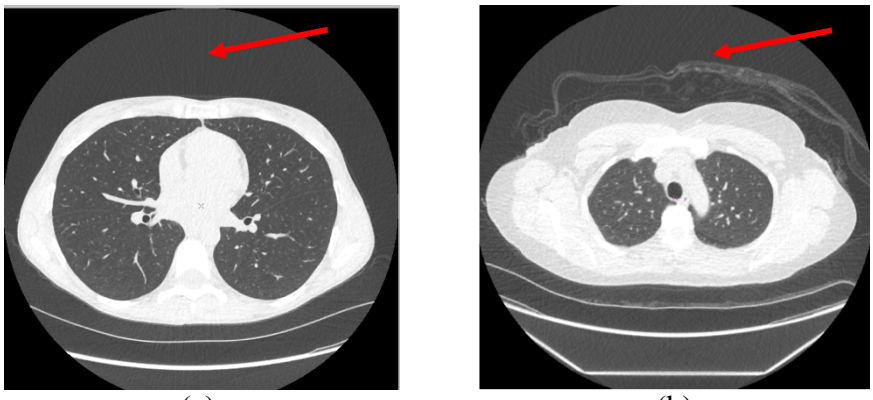
4.1 Extraction of quality measurement regions

The image quality measurement described in this dissertation includes the measurement of noise and calibration. They are both obtained from 3D homogeneous regions. Noise is represented as the pixel standard deviation of a homogeneous region and calibration is represented as the mean pixel intensity of a homogeneous region.

Previous studies in other research groups have measured noise or calibration using manually defined regions such as manually drawn circular regions inside the DA and the trachea lumen (Singh et al. 2011, Ohno et al. 2012, Yamada et al. 2012) [41-43]. In this dissertation, the measurement regions are obtained based on pre-segmented regions such as the aorta and the trachea. In total, three homogeneous regions are obtained for the purpose of quality measurement: an external air region (EA), the trachea lumen region (TA), and the DA lumen blood region (AB). The extractions of the 3D homogeneous regions are described below.

4.1.1 Extraction of the EA region

The external air region is defined to be anterior to and outside the body. It lies within the image reconstruction circle with the same vertical extent of the lungs. This is a very homogeneous air region in majority of the cases. However, occasionally artifacts such as clothing can affect the homogeneity in this region (see Figure 4.1). To avoid this issue, the algorithm described in Algorithm 4.1 is used to extract a maximized homogeneous EA region.

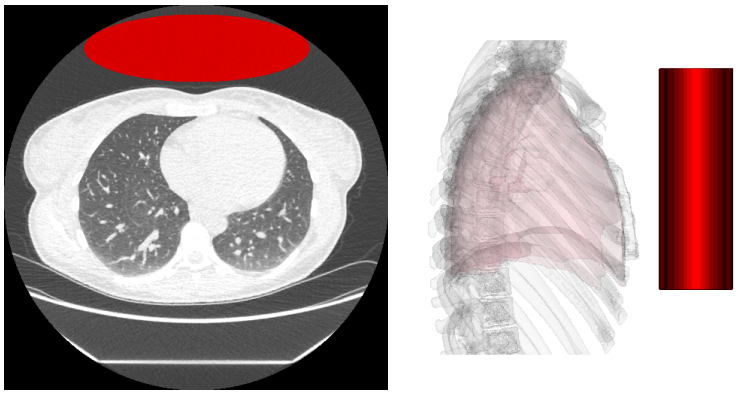


(a) (b)

Figure 4.1: Example EA regions in two cases (indicated by the red arrows). (a) a homogeneous EA region; (b) EA region with artifacts.

Algorithm 4.1 Extraction of EA region

1) Identify air pixels in the pre-segmented external body region through median image filtering and intensity thresholding at -600HU;
2) Fit the largest possible elliptical cylinder into this air region (see Figure 4.2);
3) If the elliptical cylinder does not meet the minimum size requirement (a cross sectional dimension of 20mm×10mm), use a fixed size ellipse (20mm×10mm) in image slice where this ellipse will fit in.

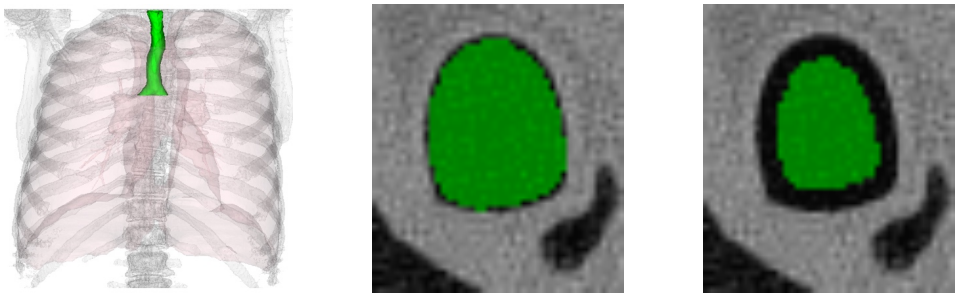


(a) (b)

Figure 4.2: Visualization of an extracted EA region. (a) EA in red in an axial slice; (b) EA in red in a sagittal view together with the lungs (light pink) and the bones (light grey).

4.1.2 Extraction of the TA region

Trachea region can be obtained from the segmented airway using a previously developed algorithm (Lee et al. 2009) [45]. Only the trachea region within the extent of the lungs is used. The trachea mask is eroded with a small kernel (circle with a diameter of 5mm) to avoid the boundary transition region. The extracted TA region is shown in Figure 4.3.

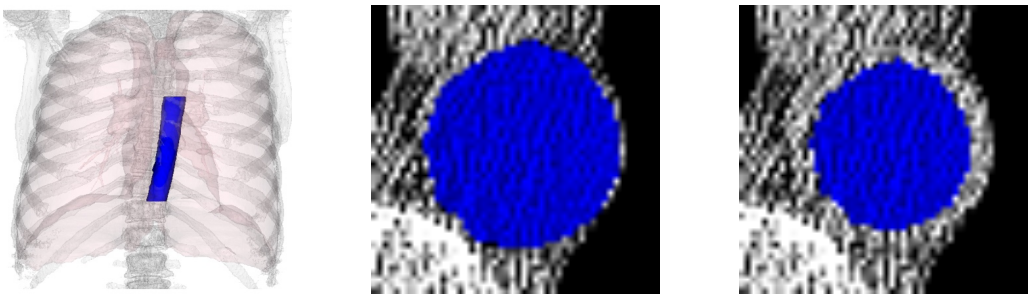


(a) (b) (c)

Figure 4.3: Visualization of an extracted TA region. (a) segmented trachea in green together with the lungs (light pink) and the bones (light grey) in a coronal view; (b) segmented trachea in green in an axial slice; (c) TA region after erosion.

4.1.3 Extraction of the AB region

The DA is obtained from the segmented aorta. It starts at the carina level and ends at the top of the lung diaphragm (i.e. the end of the segmented DA). The DA is also eroded with a small kernel (circle with a diameter of 7mm) to avoid boundary transition region. The extracted DA lumen region AB is shown in Figure 4.4.



(a) (b) (c)

Figure 4.4: Visualization of an extracted AB region. (a) segmented DA in blue together with the lungs (light pink) and the bones (light grey) in a coronal view; (b) segmented DA in blue in an axial slice; (c) AB region after erosion.

4.1.4 Quality measurement using EA, TA, and AB

Noise and calibration are measured from the three extracted homogeneous regions in two different ways. First, the global noise and calibration values are reported, i.e. the standard deviation and mean of all pixels within the EA, TA and AB regions (see Figure 4.5). Second, the noise and calibration values are reported in each image slice, resulting in a quality measurement profile as shown in Figure 4.6.

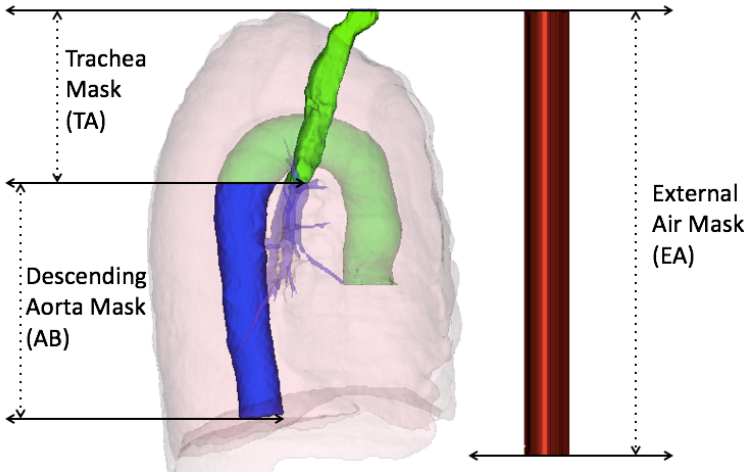
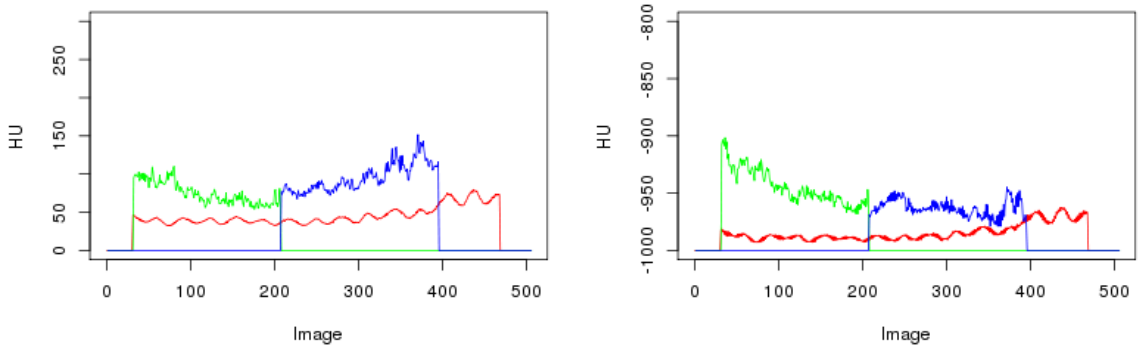


Figure 4.5: Three homogeneous regions shown in a sagittal view. EA (red), TA (green), and AB (blue) are shown together with the lungs (light pink). For visualization purpose, TA and AB are not eroded.



(a) (b)

Figure 4.6: Noise and calibration profiles along the image slices. (a) noise profiles where EA profile is in red, TA in green and AB in blue; (b) calibration profiles with the same color schemes. For plotting purpose, 1000HU has been deducted from AB calibration values.

4.2 Evaluation methods and datasets

The automated algorithm is first evaluated on phantom CT datasets to assess its repeatability and sensitivity. Then it is evaluated on a cohort of LCS LDCT) scans. Section 4.2.1 defines the experiments using the phantom data and section 4.2.2 defines the experiments using the LDCT data. All LDCT scans selected for evaluation have acceptable cardiac and airway segmentations using visual inspection to ensure the quality of the AB and TA measurement. For EA segmentation, the mask region spans a minimum of 10 image slices.

The primary image quality measurements are: noise and calibration in the EA, TA, AB

regions. Noise in the three regions is denoted as En, Tn, and An. Calibration in the three regions is denoted as Ec, Tc, and Ac. For a group of scans, statistics of the group measurement is used for quality characterization. For instance, the mean and standard deviation of En from all scans in the group is denoted as μ_{En} and σ_{En}, which are used to characterize the EA noise properties of the group.

In addition to the global noise and calibration measurements, the per-slice image quality profile is also employed to analyze groups of CT scans. For each group, an average quality profile is computed by performing a linear interpolation on the per-slice measurement from each individual scan. The profiles are interpolated to the same length (1000 sample points) and averaged to obtain the average quality profile. The standard deviation at each interpolated point is also computed and plotted together with the average.

4.2.1 Phantom experiments

The image quality measurements are performed on CT scans acquired using a chest anthropomorphic phantom [88-90]. The phantom was used in an FDA study to evaluate lung nodule size estimation and contains a vasculature insert on which synthetic nodules were inserted or attached. Each unique positioning of the nodules is referred to as a nodule layout. It also contains a trachea and bone structures. It does not contain an aorta.

The phantom was scanned using a Philips 16-row scanner (Mx8000 IDT) and a Siemens 64-row scanner (Somatom Definition 64). Scans were acquired with varying combinations of effective dose (mAs), pitch, and slice collimation, and reconstructed with varying combinations of slice thicknesses and reconstruction kernels. Ten repeat scans have been collected for each imaging protocol (acquisition and reconstruction) and each nodule layout. For the 10 repeat scans, the phantom position was not changed; however, it was repositioned between different

imaging protocols or different nodule layouts.

In the current public dataset released for this study, there are 4 datasets corresponding to 4 different nodule layouts. Three of the datasets were acquired using Philips and the fourth acquired using Siemens. The Philips scans had the following imaging protocols: four different effective dose (20 or 25mAs, 50mAs, 100mAs, 200mAs), slice collimation 16×0.75mm, slice overlap of 50%, two different pitches (0.9 and 1.2), two different reconstruction filters (B and C), multiple slice thicknesses (0.8mm, 1.5mm, 2mm, 3mm, 5mm). The corresponding z image resolution (zres) is 0.4mm, 0.7mm, 1mm, 1.5mm, 2.5mm. The Siemens scans had the following imaging protocols: three different effective dose (25mAs, 100mAs, 200mAs), slice collimation 64×0.6mm, two different overlaps (0% and 50%), two different pitches (0.9 and 1.2), two different reconstruction filters (B40f and B60f), multiple slice thicknesses (0.75mm, 1.5mm, 3mm). The corresponding zres is 0.4mm, 0.7mm, 1.5mm (with overlap) and 0.7mm, 1.5mm, 3mm (no overlap).

Our automated algorithms were able to reliably segment the EA and TA regions in most phantom images. Since DA is not represented in the phantom, image quality is only measured in EA and TA regions using cases where these two regions have been correctly segmented. Four different experiments are performed. The first experiment evaluates the algorithms' repeatability between longitudinal scan pairs with the same imaging protocol. The next three experiments evaluate the sensitivity of the algorithms to the change of imaging protocols including the radiation exposure levels, slice thicknesses and reconstruction filters; The sensitivity experiments are evaluated separately on a set of Philips scans and a set of Siemens scans. Table 4.1-4.4 list the dataset description for the four experiments.

The first experiment studies the repeatability of the measurements between longitudinal

scan pairs with the same imaging protocols. In this study, 26 longitudinal scan pairs are selected. For each pair, the two scans have the same imaging protocol but were taken at different times (T1 and T2). The time interval between T1 and T2 ranges from 4 to 7 months. The imaging protocol of each pair is summarized in Table 4.1. The hypothesis is that the algorithm should have minimum differences in all measurements between longitudinal scan pairs.

For this experiment, the statistics of the differences in each quality measurement is reported. There are 4 differences: differences in EA noise ΔEn, TA noise ΔTn, EA calibration ΔEc, and TA calibration ΔTc. Each difference is computed by subtracting T2 measurement from T1 measurement. The minimum, maximum, mean, median and standard deviation of the 26 differences are reported.

Table 4.1. Datasets summary for the longitudinal scan pairs in phantom repeatability experiment.

pair	slice (mm)	zres (mm)	mAs	pitch	collimation	recon
1	0.8	0.4	200	0.9	0.75	C
2	1.5	0.7	200	0.9	0.75	C
3	3	1.5	200	0.9	0.75	C
4	2	1	25	0.9	1.5	C
5	3	1.5	25	0.9	1.5	C
6	5	2.5	25	0.9	1.5	C
7	1.5	0.7	100	0.9	0.75	C
8	3	1.5	100	0.9	0.75	C
9	2	1	100	0.9	1.5	C
10	3	1.5	100	0.9	1.5	C
11	5	2.5	100	0.9	1.5	C
12	2	1	200	0.9	1.5	C
13	3	1.5	200	0.9	1.5	C
14	5	2.5	200	0.9	1.5	C
15	2	1	200	1.2	1.5	C
16	3	1.5	200	1.2	1.5	C
17	5	2.5	200	1.2	1.5	C
18	0.8	0.4	25	0.9	0.75	C
19	3	1.5	25	0.9	0.75	C
20	1.5	0.7	200	1.2	0.75	C
21	3	1.5	200	1.2	0.75	C
22	1.5	0.7	100	1.2	0.75	C
23	3	1.5	100	1.2	0.75	C
24	2	1	100	1.2	1.5	C
25	3	1.5	100	1.2	1.5	C
26	5	2.5	100	1.2	1.5	C

The second experiment compares the quality measurement between scans of the same imaging protocol except for radiation exposure level (mAs). The hypothesis is that a higher mAs will lead to a decreased noise level while the variations of mAs should have a minimum impact on calibration. In this experiment, for each unique combination of imaging protocols, statistics of the quality measurements are given for groups of repeated scans. In addition, for the lowest and highest radiation groups, the average quality profiles along the image slices are also computed. Dataset for this experiment is summarized in Table 4.2.

Table 4.2. Datasets summary for phantom sensitivity experiment: change of radiation exposure (mAs).

Scanner	Number of scans (exposure)	Other parameters
Philips	Group1: 10 (25mAs) Group2: 10 (50mAs) Group3: 10 (100mAs) Group4: 4 (200mAs)	Slice thickness=0.8mm zres=0.4mm Pitch=0.9 Collimation=0.75mm Recon filter=C
Siemens	Group1: 10 (25mAs) Group2: 10 (100mAs) Group3: 10 (200mAs)	Slice thickness=0.75mm zres=0.7mm Pitch=0.9 Collimation=0.6mm Recon filter=B40f

The third experiment compares the quality measurement between scans of the same imaging protocol except for reconstruction slice thicknesses. The hypothesis is that a thicker slice thickness will lead to a decreased noise level while the variations of slice thicknesses should have a minimum impact on calibration. In this experiment, for each unique combination of imaging protocols, statistics of the quality measurements are given for groups of repeated scans. Dataset for this experiment is summarized in Table 4.3.

Table 4.3. Datasets summary for phantom sensitivity experiment: change of slice thickness.

Scanner	Number of scans (slice thickness, zres)	Other parameters
Philips	Group1: 9 (0.8mm, 0.4mm) Group2: 10 (1.5mm, 0.7mm) Group3: 9 (3mm, 1.5mm)	Effective dose=25mAs Pitch=0.9 Collimation=0.75mm Recon filter=C
Siemens	Group1: 10 (0.75mm, 0.4mm) Group2: 9 (1.5mm, 0.7mm) Group3: 10 (3mm, 1.5mm)	Effective dose=25mAs Pitch=0.9 Collimation=0.6mm Recon filter=B40f

The fourth experiment compares the quality measurement between scans of the same imaging protocol except for reconstruction filter type. The hypothesis is that a more enhanced reconstruction filter will lead to an increased noise level while the variations of reconstruction filters should have a minimum impact on calibration. In this experiment, for each unique

combination of imaging protocols, statistics of the quality measurements are given for groups of repeated scans. Dataset for this experiment is summarized in Table 4.4.

Table 4.4. Datasets summary for phantom sensitivity experiment: change of reconstruction filter.

Scanner	Number of scans (recon filter)	Other parameters
Philips	Group1: 5 (B) Group2: 10 (C)	Slice thickness=2mm zres=1mm Effective dose=25mAs Pitch=0.9 Collimation=1.5mm
Siemens	Group1: 10 (B40f) Group2: 9 (B60f)	Slice thickness=0.75mm zres=0.7mm Effective dose=25mAs Pitch=0.9 Collimation=0.6mm

4.2.2 LDCT experiments

Studies from a LCS cohort [73] were used retrospectively to evaluate the repeatability of the quality measures and the impact on quality of different CT parameter settings. Four experiments are performed to assess the repeatability and sensitivity of the measurements. Description of the datasets and experiments are given below.

The first experiment evaluates the algorithm repeatability using longitudinal scan pairs with consistent imaging protocols. It is evaluated on 31 pairs of longitudinal scans. All scans in this experiment were acquired using Siemens Somatom Definition Flash scanner model and reconstructed using I41f\4 filter and at a slice thickness of 0.5mm without overlapping. They were acquired with a kVp of 120 and an mA of 48. The time interval between the pairs ranges from 84 to 438 days (median 369 days).

For this experiment, the statistics of the differences in each quality measurement is reported. There are 6 differences: differences in EA noise ΔEn, TA noise ΔTn, AB noise ΔAn, EA calibration ΔEc, TA calibration ΔTc, and AB calibration ΔAc. Each difference is computed

by subtracting T2 measurement from T1 measurement. The minimum, maximum, mean, median and standard deviation of the 31 differences are reported.

The second experiment compares the quality measurement between scans of the same imaging protocol except for reconstruction filter type. It follows the same hypothesis of the phantom experiment three (a more enhanced reconstruction filter leads to an increased noise level while having a minimum impact on calibration). However, more variations are expected compared to the phantom experiment. In this experiment, for each unique combination of imaging protocols, statistics of the quality measurements are given for groups of scans with the same reconstruction filter. The datasets are listed in Table 4.5.

Table 4.5. Datasets summary for LDCT sensitivity experiment: change of reconstruction filter.

Scanner	Number of scans (recon filter)	Other parameters
GE (Medical Systems Discovery CT750 HD)	Group1: 74 (Standard) Group2: 74 (Bone)	Slice thickness=0.625mm zres=0.625mm Exposure=120kVp, 30mA
Siemens (Biograph 40)	Group1: 160 (B40f) Group2: 160 (B60f)	Slice thickness=1mm zres=1mm Exposure=120kVp, 48mA

The third experiment also compares the quality measurements between different reconstruction filter types. The average quality profiles for each reconstruction filter type are plotted. This experiment is evaluated on four datasets, two of which have also been used in the first experiment (Siemens B40f and B60f). The other two datasets are: a group of scans acquired using dose-modulation and reconstructed with the iterative reconstruction filter; a group of scans reconstructed with the iterative reconstruction filter without dose-modulation. The hypothesis is that the average quality profiles will have different trends for all four reconstruction filters. Details of the four datasets are given in Table 4.6.

Table 4.6. Datasets summary for LDCT sensitivity experiment: change of reconstruction filter using image profiles.

Scanner + Recon Filter	Number of scans	Other parameters
Siemens (Biograph 40) + B40f	160	Slice thickness=1mm zres=1mm Exposure=120kVp, 48mA
Siemens (Biograph 60) + B60f	160	Slice thickness=1mm zres=1mm Exposure=120kVp, 48mA
Siemens (Somatom Definition Flash) + I41f\4 + dose-modulation	9	Slice thickness=0.5mm zres=0.5mm Exposure=120kVp, various mA
Siemens (Somatom Definition Flash) + I41f\4	160	Slice thickness=0.5mm zres=0.5mm Exposure=120kVp, 48mA

The fourth experiment evaluates the general statistics on quality measurements on a large LCS dataset with 841 scans. They consist of 544 GE scans (Discovery CT750 HD and LightSpeed VCT, 120kVp, 30 or 40 mA) with Standard reconstruction filters and 297 Siemens scans (Biograph40 and Sensation Cardiac 64, 120kVp, 48 or 60 mA) with B40f reconstruction filters. All scans have a slice thickness of 1.25mm or less. Histograms of the six quality measurements (En, Tn, An, Ec, Tc, Ac) are plotted.

4.3 Results

Results of the phantom and LDCT experiments are reported in this section: section 4.3.1 reports the phantom experiment results and section 4.3.2 reports the LDCT experiment results.

4.3.1 Phantom experiment results

Results of the first phantom experiment (repeatability between longitudinal scan pairs) are summarized in Table 4.7. Results of the second phantom experiment (change in radiation exposure) are given in Table 4.8 and Figures 4.7-4.9. Figures 4.7-4.8 show the average quality profiles for the lowest and highest radiation exposure settings. Results of the third and fourth phantom experiments (change in slice thickness and change in reconstruction filter) are given in

Tables 4.9-4.10. Figure 4.9 summarizes the phantom sensitivity experiment results in Tables 4.8-4.10.

Table 4.7. Phantom repeatability experiment results: statistics of the differences (HU) in image quality between longitudinal scan pairs.

	min	max	mean	sdev	median
ΔEn	-1.03	0.04	-0.18	0.25	-0.11
ΔTn	-2.18	1.24	-0.45	1.02	-0.54
ΔEc	-1.40	0.09	-0.22	0.36	-0.13
ΔTc	-3.54	1.55	0.46	1.18	0.84

Table 4.8. Phantom sensitivity experiment results: change in radiation exposure.

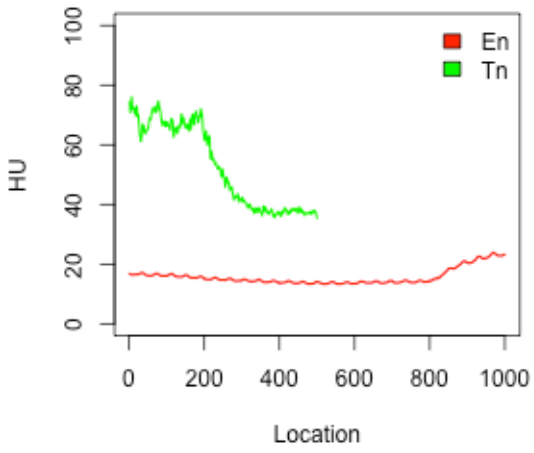
Scanner	mAs	μ_{En} (σ_{En})	μ_{Tn} (σ_{Tn})	μ_{Ec} (σ_{Ec})	μ_{Tc} (σ_{Tc})
Philip	25	16.11 (0.05)	55.39 (0.47)	-988.80 (0.04)	-887.47 (1.41)
	50	11.45 (0.03)	45.89 (0.74)	-991.70 (0.03)	-875.97 (0.70)
	100	8.34 (0.02)	34.35 (0.44)	-993.78 (0.01)	-878.30 (0.53)
	200	6.21 (0.01)	25.30 (0.15)	-995.22 (0.01)	-878.86 (0.28)
Siemens	25	19.43 (0.04)	54.39 (1.67)	-995.73 (0.02)	-894.85 (0.61)
	100	11.46 (0.03)	29.07 (0.14)	-999.71 (0.02)	-896.34 (0.32)
	200	8.71 (0.02)	22.01 (0.14)	-1000.16 (0.03)	-895.76 (0.19)

Table 4.9. Phantom sensitivity experiment results: change in slice thickness.

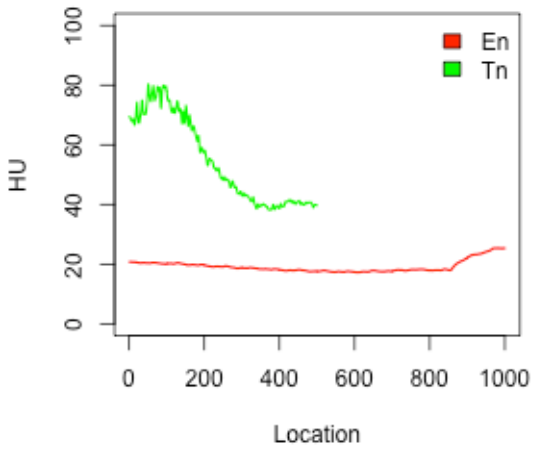
Scanner	zres	μ_{En} (σ_{En})	μ_{Tn} (σ_{Tn})	μ_{Ec} (σ_{Ec})	μ_{Tc} (σ_{Tc})
Philips	0.4	16.16 (0.04)	55.16 (0.56)	-988.81 (0.04)	-887.22 (1.72)
	0.7	11.00 (0.03)	42.79 (0.68)	-992.20 (0.03)	-889.40 (1.27)
	1.5	8.12 (0.03)	35.37 (0.74)	-994.09 (0.02)	-891.39 (2.55)
Siemens	0.4	19.44 (0.04)	54.37 (0.95)	-995.72 (0.02)	-894.79 (0.54)
	0.7	14.28 (0.03)	38.39 (0.98)	-998.26 (0.02)	-895.42 (0.42)
	1.5	11.05 (0.02)	29.41 (0.83)	-999.71 (0.02)	-895.45 (0.45)

Table 4.10. Phantom sensitivity experiment results: change in reconstruction filter.

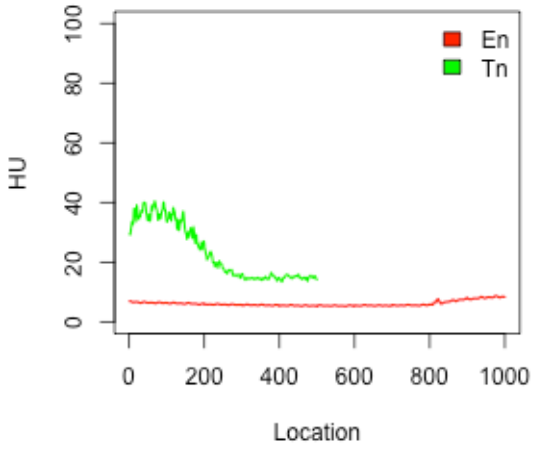
Scanner	Recon	μ_{En} (σ_{En})	μ_{Tn} (σ_{Tn})	μ_{Ec} (σ_{Ec})	μ_{Tc} (σ_{Tc})
Philips	B	7.07 (0.02)	30.81 (0.51)	-995.31 (0.02)	-874.05 (0.62)
	C	9.19 (0.03)	40.39 (0.76)	-993.92 (0.01)	-874.27 (0.95)
Siemens	B40f	19.43 (0.04)	54.39 (1.67)	-995.73 (0.02)	-894.85 (0.61)
	B60f	78.77 (0.16)	166.30 (2.87)	-949.78 (0.06)	-833.10 (2.79)



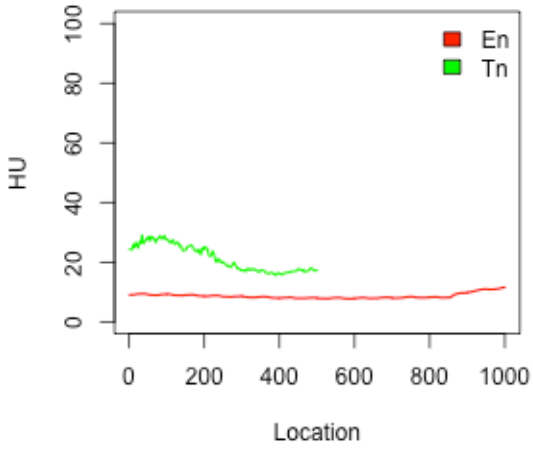
(a) Philip noise (25mAs)



(b) Siemens noise (25mAs)

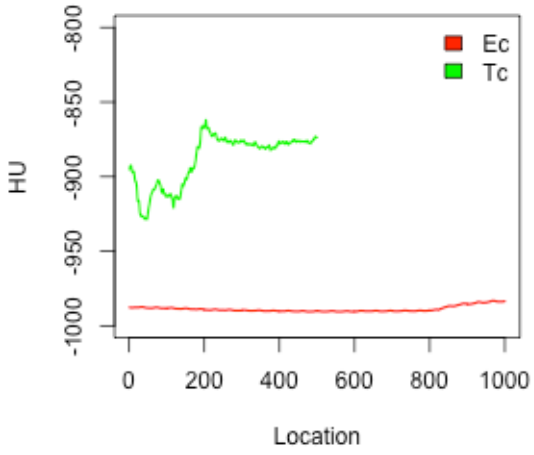


(c) Philip noise (200mAs)

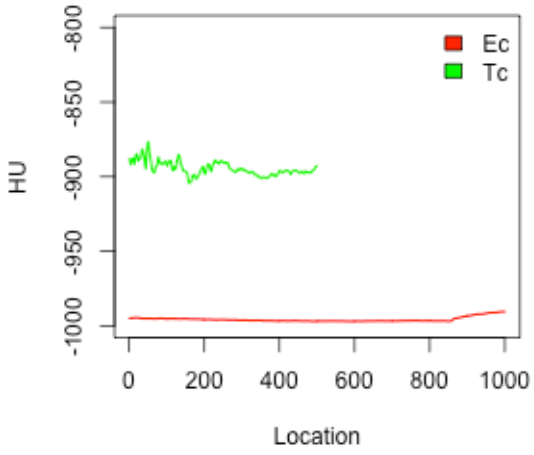


(d) Siemens noise (200mAs)

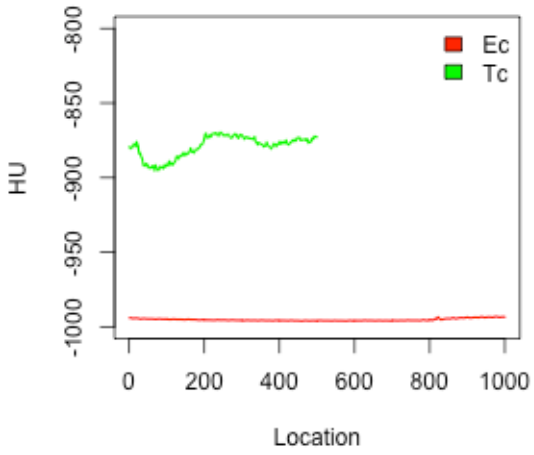
Figure 4.7: Average noise profiles from phantom sensitivity experiment: change of mAs (lowest and highest mAs shown). (a) Philip cohort (25mAs); (b) Siemens cohort (25mAs); (c) Philip cohort (200mAs); (d) Siemens cohort (200mAs).



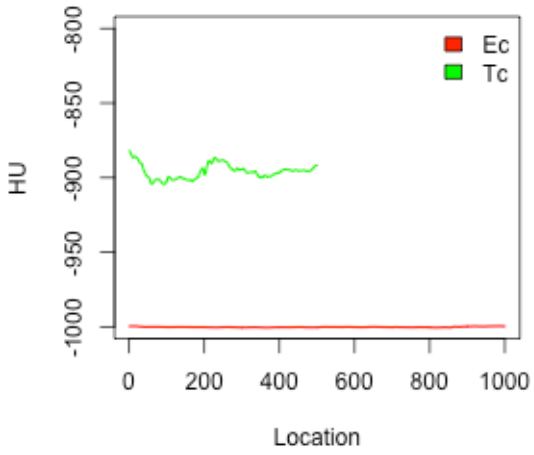
(a) Philip calibration (25mAs)



(b) Siemens calibration (25mAs)

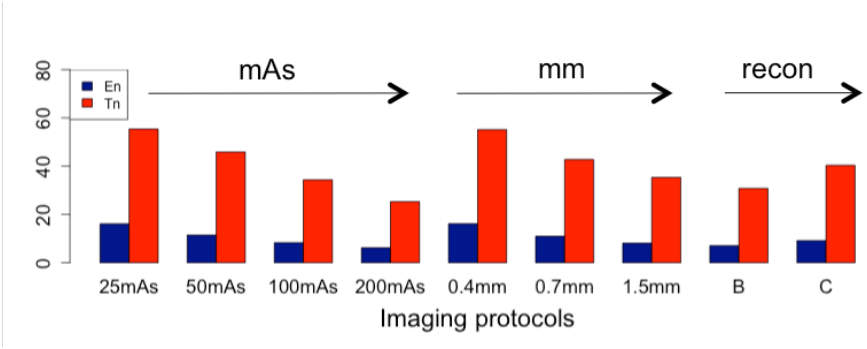


(c) Philip calibration (200mAs)

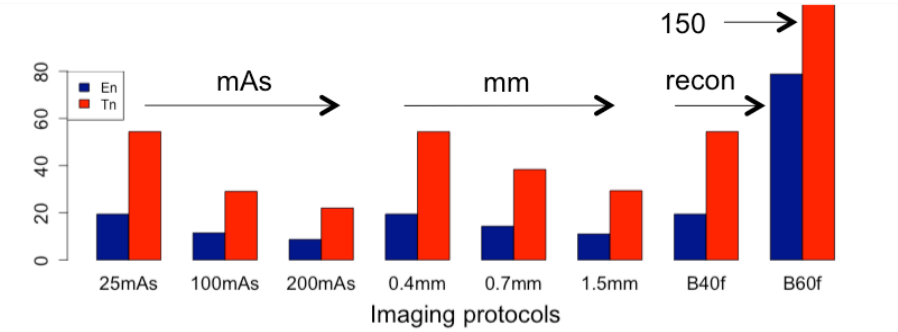


(d) Siemens calibration (200mAs)

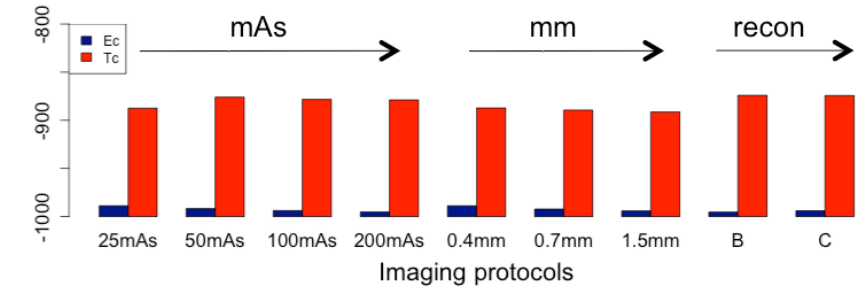
Figure 4.8: Average calibration profiles from phantom sensitivity experiment: change of mAs (lowest and highest mAs shown). (a) Philip cohort (25mAs); (b) Siemens cohort (25mAs); (c) Philip cohort (200mAs); (d) Siemens cohort (200mAs).



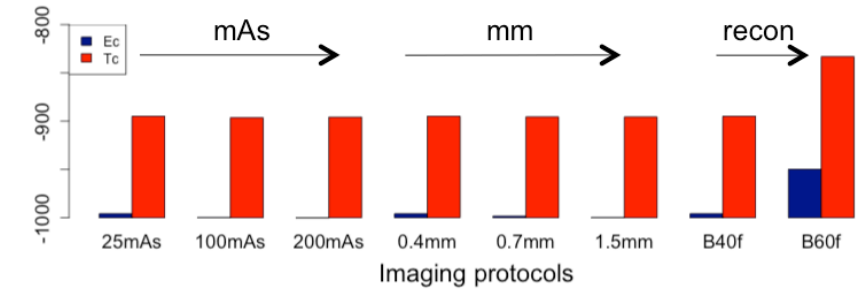
(a) Philips noise



(b) Siemens noise



(c) Philips calibration



(d) Siemens calibration

Figure 4.9: Summary plot for phantom sensitivity experiments. (a) Philips cohort noise; (b) Siemens cohort noise; (c) Philips cohort calibration; (d) Siemens cohort calibration.

4.3.2 LDCT experiment results

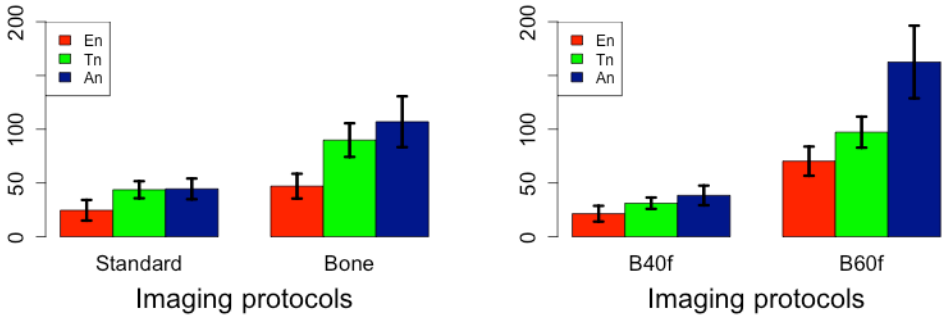
Results of the first LDCT experiment (repeatability between longitudinal scan pairs) are summarized in Table 4.11. Results of the second and third LDCT experiment (change in reconstruction filter) are given in Table 4.12 and Figures 4.10-4.12. Figure 4.10 shows the statistics for each imaging protocol. The plots are grouped together based on scanner types. Figures 4.11-4.12 show the average quality profiles for the four datasets. Results of the fourth LDCT experiment (histogram statistics on a large LCS cohort) are shown in Figure 4.13 and Table 4.13.

Table 4.11. LDCT repeatability experiment results: statistics of the differences (HU) in image quality between longitudinal scan pairs.

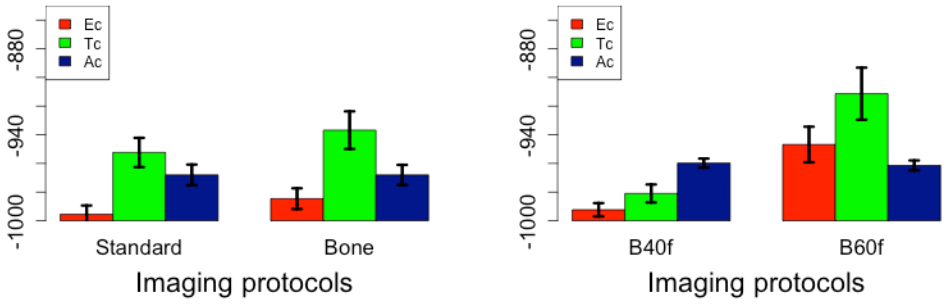
	min	max	mean	sdev	median
ΔEn	-48.44	9.28	-4.40	11.18	-1.42
ΔTn	-5.29	4.44	-1.62	2.58	-2.25
ΔAn	-5.48	4.08	-1.01	2.59	-0.81
ΔEc	-31.03	17.12	0.21	9.55	1.74
ΔTc	-9.82	30.05	4.92	9.58	2.16
ΔAc	-7.85	5.16	-1.27	2.73	-0.81

Table 4.12. LDCT sensitivity experiment results: change in reconstruction filter.

Scanner	GE		Siemens	
Recon	Standard	Bone	B40f	B60f
μ_{En} (σ_{En})	24.61 (9.60)	47.07 (11.52)	21.44 (7.32)	70.32 (13.57)
μ_{Tn} (σ_{Tn})	43.74 (7.94)	89.91 (15.65)	31.20 (5.27)	97.28 (14.42)
μ_{An} (σ_{An})	44.53 (9.64)	106.99 (23.64)	38.45 (9.13)	162.58 (33.76)
μ_{Ec} (σ_{Ec})	-995.53 (6.17)	-984.60 (7.26)	-992.33 (4.64)	-946.81 (12.41)
μ_{Tc} (σ_{Tc})	-952.36 (10.19)	-936.78 (13.11)	-980.97 (6.27)	-911.37 (18.17)
μ_{Ac} (σ_{Ac})	32.00 (7.23)	31.99 (7.00)	40.22 (3.15)	38.63 (3.45)

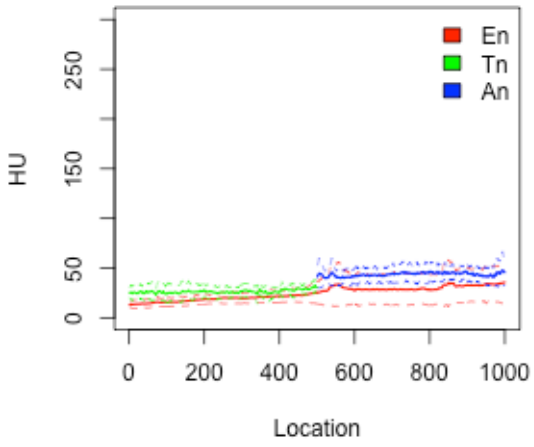


(a) GE noise (b) Siemens noise

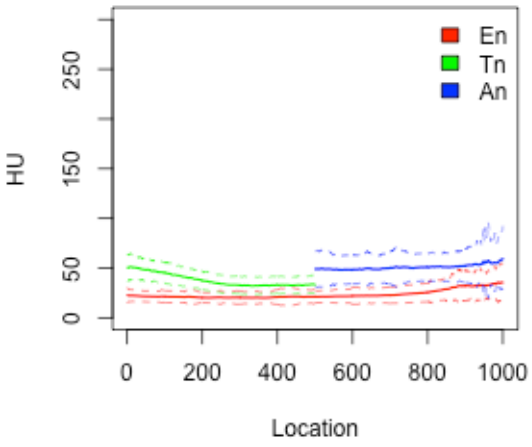


(c) GE calibration (d) Siemens calibration

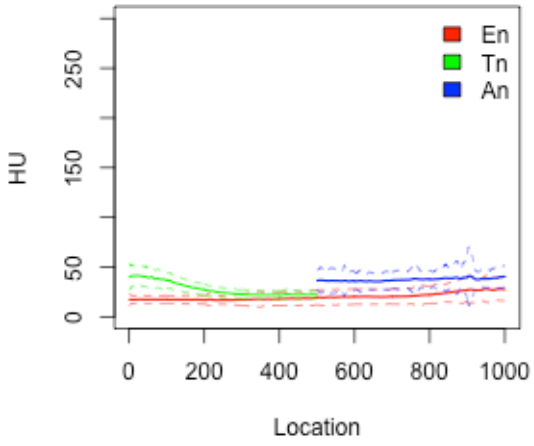
Figure 4.10: Summary bar plot showing mean and standard deviation for the LDCT sensitivity experiment. (a) GE cohort noise; (b) Siemens cohort noise; (c) GE cohort calibration; (d) Siemens cohort calibration.



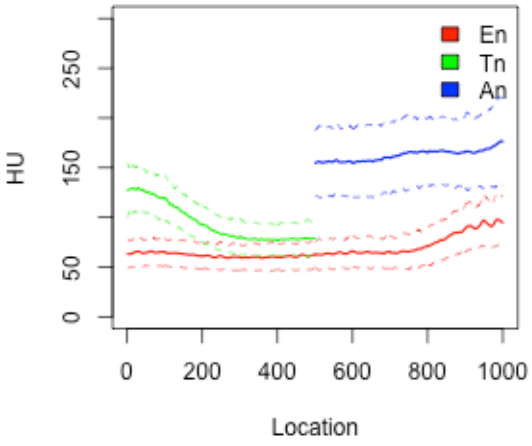
(a) Dose-modulated I41f\4



(b) Regular I41f\4

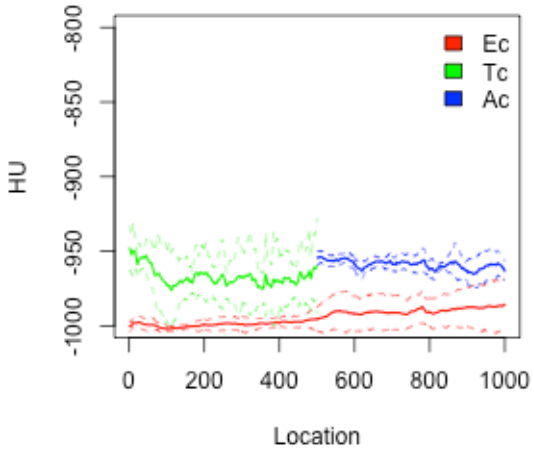


(c) B40f

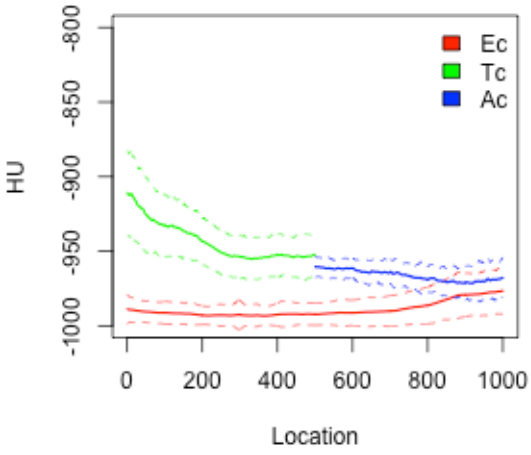


(d) B60f

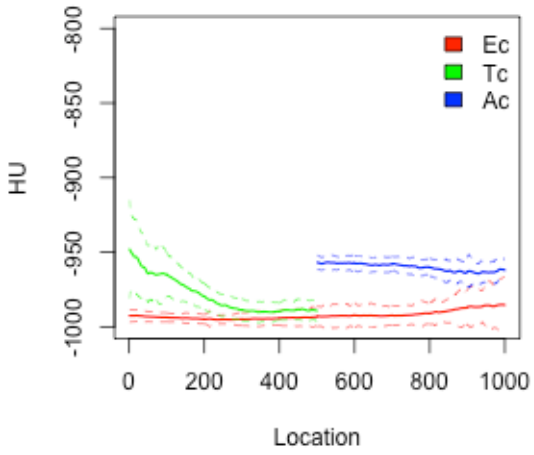
Figure 4.11: Average noise profiles from LDCT sensitivity experiment.



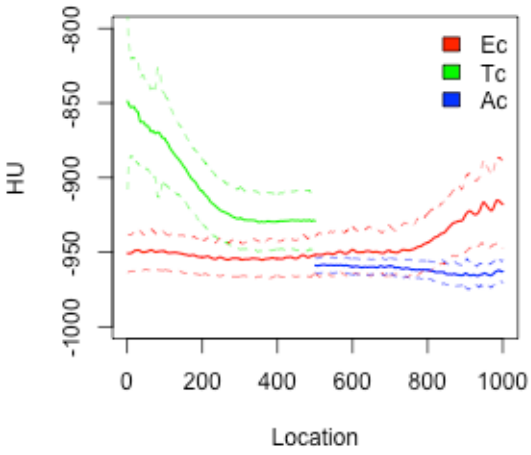
(a) Dose-modulated I41f\4



(b) Regular I41f\4

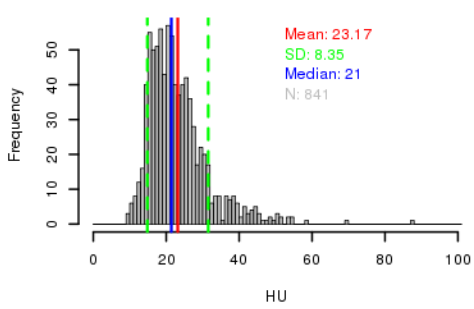


(c) B40f

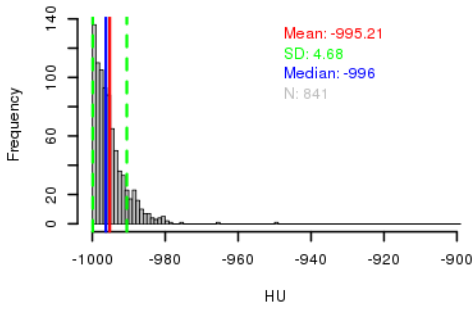


(d) B60f

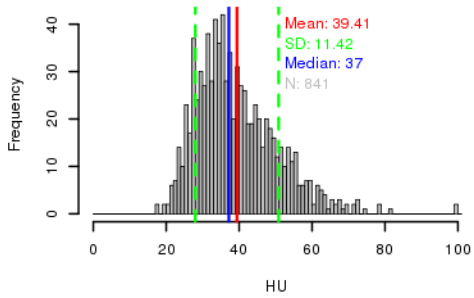
Figure 4.12: Average calibration profiles from LDCT sensitivity experiment.



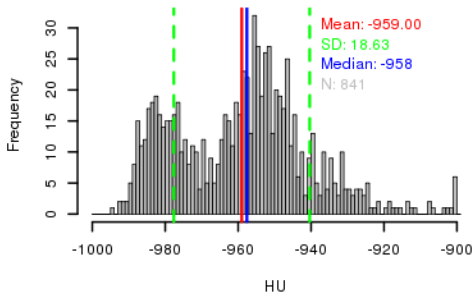
(a) En



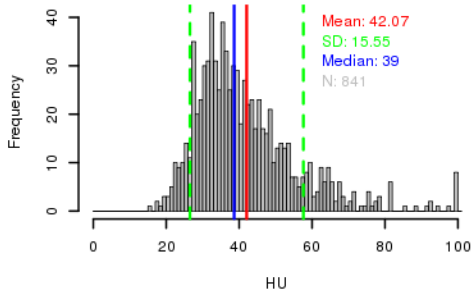
(b) Ec



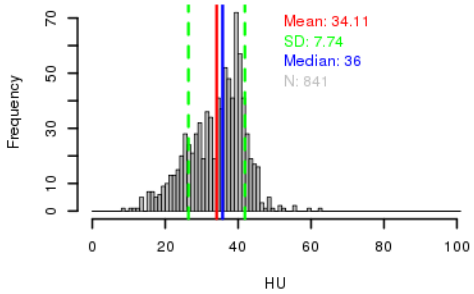
(c) Tn



(d) Tc



(e) An



(f) Ac

Figure 4.13: Noise and calibration histograms for the LCS cohort. (a) En; (b) Ec; (c) Tn; (d) Tc; (e) An; (f) Ac.

Table 4.13. LDCT experiment results: statistics of quality on a LCS cohort (841 scans).

Measure	Min	Max	Mean	Standard deviation
En	9.22	87.40	23.17	8.35
Tn	17.60	118.60	39.41	11.42
An	15.45	171.50	42.07	15.55
Ec	-1006.0	-949.6	-995.2	4.68
Tc	-994.8	-881.3	-959.0	18.63
Ac	8.50	62.53	34.11	7.74

4.4 Discussion

The automated CT image quality assessment algorithm has shown good repeatability and sensitivity in both phantom and LDCT experiments. In the phantom repeatability experiment, the largest standard deviation among all 4 types of differences (ΔEn, ΔTn, ΔEc, ΔTc) is 1.18HU as shown in Table 4.7, which indicates that under ideal situations, these quality measurements are very repeatable. In the LDCT repeatability experiment, the largest standard deviation among all 6 types of differences (ΔEn, ΔTn, ΔAn, ΔEc, ΔTc, ΔAc) is 11.18HU as shown in Table 4.11, larger than that from the phantom experiment. It should also be noted that the AB repeatability is the highest among the three regions in the LDCT experiment (standard deviation of 2.59HU for ΔAn and 2.73HU for ΔAc).

In the phantom sensitivity experiment, the noise measurement has exhibited expected changes regarding the changes in imaging protocols. For instance, when radiation exposure or slice thickness increases, the noise level in both EA and TA decreases; images reconstructed with a more enhanced filter type (B60f, C) have higher noise than those reconstructed with a less enhanced filter (B40f, B). The noise measurement has displayed similar characteristics in the LDCT sensitivity experiment, i.e. the Bone reconstruction filter corresponds to a higher noise level than the Standard filter and the B60f filter higher than the B40f filter.

In the phantom experiments, the calibration measurement has displayed a small change (in the order of 5HU) regarding the changes in radiation exposure and slice thickness. The calibration measurement has a small change between reconstruction filter B and C (less than 5HU) but has a larger change between reconstruction filter B40f and B60f (in the order of 50HU). This characteristic has also been displayed in the LDCT experiment: calibration in the air regions has a large change between B40f and B60f (in the order of 50HU) and a relatively

large change between Standard and Bone (in the order of 20HU). However, calibration in AB is more stable and the change between reconstruction filters is around 2HU.

The averaged group quality profile differs with imaging protocols. In the phantom experiment, the most significant difference occurs in the upper trachea region. In the LDCT experiment, the four different reconstruction filters correspond to distinctively different profiles. The profiles from the dose-modulated scans appear the most stable and have small variations across spatial locations. The profiles from B60f reconstructed scans have the largest variations for different spatial locations: the upper TA region has higher noise and higher calibration values compared to the lower TA region; the lower EA region has higher noise and higher calibration values compared to the upper EA region; however, AB region appears to be more stable than EA and TA. The same trend has been observed in B40f and I41f\4 (without dose-modulation) reconstructed scans to a lesser extent.

When evaluated on a large LCS cohort with LDCT scans, both noise and calibration have displayed some variation among the different scans. For noise, AB region noise has both the largest variation (standard deviation of 15.55HU) and the highest mean value (42.07HU) while EA region has the lowest noise (mean = 23.17HU and standard deviation = 8.35HU). For calibration, TA region has shown the largest variation among the three regions (standard deviation of 18.63HU). It should also be noted that all histograms have exhibited a Gaussian like distribution except for the Tc measurement, which is likely caused by mixing the GE and Siemens scans together since they have very different overall Tc values. Therefore, in the setting of LCS, the automated quality measurements can be used to adjust imaging protocols by providing feedbacks of scan quality.

4.5 Conclusion

In conclusion, a fully-automated algorithm has been developed to measure CT image quality including the noise and calibration level from homogeneous 3D regions automatically extracted from CT scans. The 3D regions are: an external air region, a trachea lumen air region, and a descending aorta lumen blood region. Both global quality and quality per image slice is computed. The algorithm has been evaluated on both phantom CT scans and LCS LDCT scans with respect to its repeatability between longitudinal scan pairs and its sensitivity to the change of imaging protocols. Based on the evaluation results, the algorithm has displayed repeatability between longitudinal scans, sensitivity to imaging protocol changes, and overall robustness. It is also able to provide distinctive quality profiles for different scanners and imaging protocols. This algorithm may be used to monitor image quality retrospectively and to aid the design of computer algorithms.

CHAPTER 5

CONCLUSIONS

Fully-automated computer methods have been developed and validated to perform quantitative image biomarker measurements in chest CT scans in this research, with a special focus on the low-dose chest CT scans associated with LCS. The regulatory implementation of LCS [1] provides the opportunity to apply these methods to a large at risk population who are receiving annual LDCT. Since they are also at risk for other diseases such as coronary heart diseases, it is beneficial to provide companion diagnostics using these LDCT scans in addition to the detection and monitoring of pulmonary nodules.

The main contributions of this research are the development and validation of computer methods for the following tasks:

1) Fully-automated measurement of cardiac biomarkers (primarily CAC) through the region segmentation of cardiac organs (aorta, heart and PT);
2) Implementation of experiments to evaluate the impact of nodule size on nodule characterization in the context of small pulmonary nodules;
3) Fully-automated detection of early-stage lung diseases including ILD and pleural thickening;
4) Fully-automated quality assessment of CT scans (noise and calibration).

All segmentation and image biomarker measurement algorithms explored are fully-automated. The image segmentation algorithms have been evaluated on large datasets with more than 20,000 CT scans.

5.1 Cardiac segmentation and disease biomarker measurement

It is beneficial to provide cardiac health assessment for LCS participants since they are

also at high risk for cardiovascular diseases. For the assessment of aorta for diameter abnormalities and calcifications, a three-stage algorithm has been developed to track and segment the aorta surface from LDCT. Then the pulmonary artery trunk and the general heart region are segmented using the constraint from the segmented aorta as well as other adjacent organs. Aorta diameter profile and PT average diameter are computed from the segmented regions. Based on the heart region mask and the other segmented cardiac organs, the CAC is segmented and labeled by the artery it is associated with. In addition, calcification on the surface of the aorta and fat around the visceral heart region is segmented.

The cardiac segmentation has been evaluated both quantitatively and qualitatively. Quantitative evaluation shows that the segmented regions have good agreement with the manually marked boundaries (DSC = 0.93 for aorta, 0.88 for PT, and 0.93 for CVF). Qualitative evaluation shows when evaluated on a large dataset with over 20,000 CT scans, the algorithms produce good or acceptable results on 92% of the scans based on visual inspections. The cardiac biomarker measurements have been compared to manual measurements such as manual markings or visual scores. The CAC scores are 0.90 correlated with scores from manual markings and 0.87 correlated with visual categorical scores. The AC scores are 0.99 correlated with scores from manual markings.

This research work has shown that it is possible to perform cardiac organ segmentation and biomarker measurement in LDCT both precisely and robustly. There are a number of future directions for the cardiac analysis. First, there is still room for algorithm improvement such as the algorithm for CAC segmentation and labeling. The current algorithm relies on the spatial locations and distances to pre-segmented organs to separate out calcification in different arteries, which sometimes mistakes or misses the small regions of calcification in the LCX artery group.

Calcification in this artery usually has a small volume and is not well-constrained spatially. Therefore, a potentially more complex model is needed to correctly capture the LCX calcification. Second, it may be possible to extend the current heart segmentation algorithm to determine the actual heart boundary as well as the individual chambers. Finally, it is important to evaluate the clinical usefulness of the various cardiac biomarker measurements in a large dataset with known patient histories or diagnostic outcomes. Currently the algorithms have shown good agreement with visual reviews by radiologists. Naturally the next step is to demonstrate their usefulness such as their abilities to capture early-stage diseases or to predict disease progressions.

5.2 Lung disease biomarker measurement

The primary purpose of annual LCS is to detect a pulmonary nodule in its early-stages when it is more curable by surgery. Therefore, the focus of nodule status characterization should be on small pulmonary nodules with a diameter much less than 30mm (LCS detected nodules are often less than 10 mm) instead of the much larger nodules observed in cancer center CT scans. However, the nodule characterization studies in the literature often do not consider the size distribution of the nodule datasets and the results reported in these studies could simply be reflecting the natural size difference between malignant and benign nodules rather than the appearance difference between a malignant and a benign nodule of the same size.

In this research, a large dataset of 736 small pulmonary nodules (diameter 3-30mm) obtained from LCS is used to evaluate the impact of nodule size on characterization outcome. From the experiments, it is concluded that: 1) without taking malignant and benign nodule size difference into consideration, the characterization result is overly optimistic and reflects the size distribution difference of the malignant and benign nodules; 2) evaluation on a size-balanced

dataset (i.e. malignant and benign nodules with the same size distribution) reflects the classifier's ability to distinguish malignant and benign nodules based on their appearance features; 3) when evaluated on a size-balanced dataset, classifiers trained on a size-balanced dataset have better performance than classifiers trained on an unbalanced dataset.

There are several future directions for this work. First, a larger dataset is needed to assess and potentially improve the algorithm. Although 736 nodules is a large number for medical image related research, it is still far too small for most modern machine learning techniques. Second, other techniques such as convolutional neural networks (CNNs) could be explored for the same nodule characterization task. Methods such as CNNs do not need the design of specific features and instead learn them from the image themselves. It would be interesting to compare the results from these methods to that from a traditional machine learning classifier with explicitly designed image features.

In addition to nodule characterization, algorithms have also been developed in this research to capture early-stage ILD and pleural thickening. The ILD is represented by a fractional high-density feature obtained from partitioned lung regions with the same volume. This feature can be used to distinguish UIP (a type of ILD) from normal cases with a highest AUC of 0.95. Pleural thickening is represented by two surrogate measurements: the distance between pleural surface and ribs and the volume of concave regions on the pleural surface. These measurements have weak to moderate correlations with visual scores (distance measurement 0.40 correlated and concave volume measurement 0.74 correlated with visual scores).

Both ILD and pleural thickening detection algorithms are still in their preliminary stages and can be refined with more data available. In the future, when more ILD cases become available covering a spectrum of different disease subtypes, the algorithm can be evaluated and

potentially revised with more features and different partitioning strategies to distinguish between different ILD subtypes, which is a more challenging topic than to distinguish between ILD and normal cases. The evaluation of pleural thickening detection algorithm is very challenging since the early-stage pleural thickening is usually not visually detectable in some cases with traditional viewing windows, making the reference visual scores less reliable. In the future, it is important to gather a pleural thickening cohort with diagnostic outcomes to properly validate the automatically detected pleural thickening regions.

5.3 Image quality measurement

CT image quality such as noise level and calibration accuracy has a direct impact on computer algorithm performance. Specific algorithm designs such as parameter settings often need to be adjusted based on the actual image noise and calibration level. For instance, the CAC segmentation algorithm described in Chapter 2 requires pre-filtering and the specification of a calcification region intensity threshold, both of which are adjusted based on image noise levels.

In this research, image noise and calibration are measured in three automatically extracted homogeneous regions: an external air region EA, a trachea lumen air region TA, and a descending aorta lumen blood region AB. For each region, a global measurement and a per slice measurement are provided. This algorithm has been evaluated on both phantom CT datasets and LCS LDCT datasets and observations are made that: 1) the algorithm is repeatable for longitudinal scans; 2) the algorithm is sensitive to imaging protocol changes such as the radiation exposure level, the reconstruction slice thickness, and the reconstruction filter; 3) the algorithm can be used to demonstrate common as well as unique quality profile traits for different scanners and imaging protocols.

This algorithm serves as a first step for automated image quality measurement and

feedback. In the future, it can be used as a guidance to the modification of algorithm parameter settings. For example, the noise measurement can be used to modify the image pre-filtering level in cardiac segmentation and lung health evaluation. The calibration measurement can be used to modify biomarker measurements that depend on an absolute threshold such as the emphysema indices and the fractional high-density measurement used for ILD characterization. It can also be used to exclude images that do not meet specific quality criteria from being further analyzed.

REFERENCES

[1] T. S. Jensen, J. Chin, L. Ashby, J. Hermansen, and J. D. Hutter, "Decision Memo for Screening for Lung Cancer with Low Dose Computed Tomography (LDCT) (CAG-00439N)," https://www.cms.gov/medicare-coverage-database/details/nca-decision-memo.aspx?NCAId=274. Accessed July 1, 2016.

[2] V. A. Moyer, on behalf of the U. S. Preventive Services Task Force, "Screening for Lung Cancer: U. S. Preventive Services Task Force Recommendation Statement," *Annals of Internal Medicine*, vol. 160, pp. 330-338, Mar. 2014.

[3] A. Jemal, and S. A. Fedewa, "Lung Cancer Screening With Low-Dose Computed Tomography in the United States-2010 to 2015," *JAMA Oncology*, Feb, 2017. DOI: 10.1001/jamaoncol.2016.6416.

[4] Centers for Disease Control and Prevention, "Heart Disease Fact Sheet," https://www.cdc.gov/dhdsp/data_statistics/fact_sheets/fs_heart_disease.htm. Accessed February. 18, 2017.

[5] American Cancer Society, "Cancer Facts & Figures 2015," https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2015/cancer-facts-and-figures-2015.pdf. Accessed June 19, 2017.

[6] The International Early Lung Cancer Action Program Investigators, "Survival of Patients with Stage I Lung Cancer Detected on CT Screening," *The New England Journal of Medicine*, vol. 355, pp. 1763-1771, Oct. 2006.

[7] S. B. Edge, and C. C. Compton, "The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM," *Annals of the Surgical Oncology*, vol. 16, pp. 1471-1474, Jun. 2010.

[8] National Cancer Institute, Surveillance, Epidemiology, and End Results Program, "Cancer Stat Fact Sheets: Lung and Bronchus Cancer," http://seer.cancer.gov/statfacts/html/lungb.html. Accessed July 1, 2016.

[9] M. T. Jaklitsch, F. L. Jacobson, J. H. M. Austin, J. K. Field, J. R. Jett, S. Keshavjee, H. MacMahon, J. L. Mulshine, R. F. Munden, R. Salgia, G. M. Strauss, S. J. Swanson, W. D. Travis, and D. J. Sugarbaker, "The American Association for Thoracic Surgery guidelines for lung cancer screening using low-dose computed tomography scans for lung cancer survivors and other high-risk groups," *The Journal of Thoracic and Cardiovascular Surgery*, vol. 144, pp. 33-

38, Jul. 2012.

[10] C. I. Henschke, "International Early Lung Cancer Action Program: Screening Protocol," http://www.ielcap.org/sites/default/files/I-ELCAP-protocol.pdf. Accessed June 19, 2017.

[11] Gray's Anatomy, "Heart and lungs," https://commons.wikimedia.org/wiki/File%3AHeart-and-lungs.jpg. Accessed June 19, 2017.

[12] G. N. Hounsfield, "Computed medical imaging," Nobel lecture, December 8, 1979, *Journal of Computer Assisted Tomography*, vol. 4, pp. 665-674, Oct. 1980.

[13] K. Ono, T. Hiraoka, A. Ono, E. Komatsu, T. Shigenaga, H. Takaki, T. Maeda, H. Ogusu, S. Yoshida, K. Fukushima, and M. Kai, "Low-dose CT scan screening for lung cancer: comparison of images and radiation doses between low-dose CT and follow-up standard diagnostic CT," *Springerplus*, vol. 2, pp. 393, Aug. 2013.

[14] V. Vardhanabhuti, R. J. Loader, G. R. Mitchell, R. D. Riordan, and C. A. Roobottom, "Image Quality Assessment of Standard- and Low-Dose Chest CT Using Filtered Back Projection, Adaptive Statistical Iterative Reconstruction, and Novel Model-Based Iterative Reconstruction Algorithms," *American Journal of Roentgenology*, vol. 200, pp. 545-552, Mar. 2013.

[15] Y. Zhao, G. H. de Bock, R. Vliegenthart, R. J. van Klaveren, Y. Wang, L. Bogoni, P. A. de Jong, W. P. Mali, P. M. A. van Ooijen, and M. Oudkerk, "Performance of computer-aided detection of pulmonary nodules in low-dose CT: comparison with double reading by nodule volume," *European Radiology*, vol. 22, pp. 2076-2084, Oct. 2012.

[16] F. Beyer, L. Zierott, E. M. Fallenberg, K. U. Juergens, J. Stoeckel, W. Heindel, and D. Wormanns, "Comparison of sensitivity and reading time for the use of computer-aided detection (CAD) of pulmonary nodules at MDCT as concurrent or second reader," *European Radiology*, vol. 17, pp. 2941-2947, May 2007.

[17] O. Ecabert, J. Peters, H. Schramm, C. Lorenz, J. von Berg, M. J. Walker, M. Vembar, M. E. Olszewski, K. Subramanyan, G. Lavi, and J. Weese, "Automatic model-based segmentation of the heart in CT images," *IEEE Transactions on Medical Imaging*, vol. 27, pp. 1189-1201, Sep. 2008.

[18] B. M. Kelm, S. Mittal, Y. Zheng, A. Tsymbal, D. Bernhardt, F. Vega-Higuera, S. K. Zhou, P. Meer, and D. Comaniciu, "Detection, Grading and Classification of Coronary Stenoses in Computed Tomography Angiography," *Medical Image Computing and Computer Assisted Intervention*, vol. 6893, pp. 25-32, Sep. 2011.

[19] J. M. Wolterink, T. Leiner, R. A. P. Takx, M. A. Viergever, and I. Isgum, "Automatic Coronary Calcium Scoring in Non-Contrast-Enhanced ECG-Triggered Cardiac CT With Ambiguity Detection," *IEEE Transactions on Medical Imaging*, vol. 34, pp. 1867-1878, Sep. 2015.

[20] J. Shemesh, C. I. Henschke, A. Farooqi, R. Yip, D. F. Yankelevitz, D. Shaham, and O. S. Miettinen, "Frequency of coronary artery calcification on low-dose computed tomography screening for lung cancer," *Clinical Imaging*, vol. 30, pp. 181-185, May-Jun. 2006.

[21] J. Shemesh, C. I. Henschke, D. Shaham, R. Yip, A. O. Farooqi, M. D. Cham, D. I. McCauley, M. Chen, J. P. Smith, D. M. Libby, M. W. Pasmantier, and D. F. Yankelevitz, "Ordinal scoring of coronary artery calcifications on low-dose CT scans of the chest is predictive of death from cardiovascular disease," *Radiology*, vol. 257, pp. 541-548, Nov. 2010.

[22] I. Isgum, M. Staring, A. Rutten, M. Prokop, M. A. Viergever, and B. van Ginneken, "Multi-atlas-based segmentation with local decision fusion--application to cardiac and aortic segmentation in CT scans," *IEEE Transactions on Medical Imaging*, vol. 28, pp. 1000-1010, Jul. 2009.

[23] I. Isgum, M. Prokop, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Automatic coronary calcium scoring in low-dose chest computed tomography," *IEEE Transactions on Medical Imaging*, vol. 31, pp. 2322-2334, Dec. 2012.

[24] S. K. Shah, M. F. McNitt-Gray, S. R. Rogers, J. G. Goldin, R. D. Suh, J. W. Sayre, I. Petkovska, H. J. Kim, and D. R. Aberle, "Computer Aided Characterization of the Solitary Pulmonary Nodule Using Volumetric and Contrast Enhancement Features," *Academic Radiology*, vol. 12, pp. 1310-1319, Oct. 2005.

[25] T. W. Way, L. M. Hadjiiski, B. Sahiner, H. P. Chan, P. N. Cascade, E. A. Kazerooni, N. Bogot, and C. Zhou, "Computer-aided diagnosis of pulmonary nodules on CT scans: Segmentation and classification using 3D active contours," *Medical Physics*, vol. 33, pp. 2323-2337, Jul. 2006.

[26] D. E. Ost, and M. K. Gould, "Decision Making in Patients with Pulmonary Nodules," *American Journal of Respiratory and Critical Care Medicine*, vol. 185, pp. 363-372, Feb. 2012.

[27] K. Suzuki, F. Li, S. Sone, and K. Doi, "Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network," *IEEE Transactions on Medical Imaging*, vol. 24, pp. 1138-1150, Sep. 2005.

[28] M. Aoyama, Q. Li, S. Katsuragawa, F. Li, S. Sone, and K. Doi, "Computerized scheme for determination of the likelihood measure of malignancy for pulmonary nodules on low-dose CT images," *Medical Physics*, vol. 30, pp. 387-394, Mar. 2003.

[29] S. G. Armato 3rd, M. B. Altman, J. Wilkie, S. Sone, F. Li, K. Doi, and A. S. Roy, "Automated lung nodule classification following automated nodule detection on CT: a serial approach," *Medical Physics*, vol. 30, pp. 1188-1197, Jun. 2003.

[30] A. Jirapatnakul, A. P. Reeves, T. V. Apanasovich, A. Biancardi, D. F. Yankelevitz, and C. I. Henschke, "Pulmonary nodule classification: size distribution issues," *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1248-1251, May 2007.

[31] J. P. Wisnivesky, D. Yankelevitz, and C. I. Henschke, "The effect of tumor size on curability of stage I non-small cell lung cancers," *Chest*, vol. 126, pp. 761-765, Sep. 2004.

[32] A. McWilliams, M. C. Tammemagi, J. R. Mayo, H. Roberts, G. Liu, K. Soghrati, K. Yasufuku, S. Martel, F. Laberge, M. Gingras, S. Atkar-Khattra, C. D. Berg, K. Evans, R. Finley, J. Yee, J. English, P. Nasute, J. Goffin, S. Puksa, L. Stewart, S. Tsai, M. R. Johnston, D. Manos, G. Nicholas, G. D. Goss, J. M. Seely, K. Amjadi, A. Tremblay, P. Burrowes, P. MacEachern, R. Bhatia, M-S Tsao, and S. Lam, "Probability of Cancer in Pulmonary Nodules Detected on First Screening CT," *The New England Journal of Medicine*, vol. 369, pp. 910-919, Sept. 2013.

[33] Mayo Clinic, "Diseases and Conditions, Interstitial lung disease," http://www.mayoclinic.org/diseases-conditions/interstitial-lung-disease/basics/definition/con-20024481. Accessed March 5, 2017.

[34] A. Depeursinge, D. Van de Ville, A. Platon, A. Geissbuhler, P. Poletti, and H. Muller, "Near-Affine-Invariant Texture Learning for Lung Tissue Analysis Using Isotropic Wavelet Frames," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, pp. 665-675, Jul. 2012.

[35] Y. Xu, E. J. van Beek, Y. Hwanjo, J. Guo, G. McLennan, E. A. Hoffman, "Computer-aided classification of interstitial lung diseases via MDCT: 3D adaptive multiple feature method (3D AMFM)," *Academic Radiology*, vol. 13, pp. 969-978, Aug. 2006.

[36] P. D. Korfiatis, A. N. Karahaliou, A. D. Kazantzi, C. Kalogeropoulou, and L. I. Costaridou, "Texture-Based Identification and Characterization of Interstitial Pneumonia Patterns in Lung Multidetector CT," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, pp. 675-680, May 2010.

[37] P. A. Gevenois, V. de Maertelaer, A. Madani, C. Winant, G. Sergent, and P. De Vuyst, "Asbestosis, pleural plaques and diffuse pleural thickening: three distinct benign responses to asbestos exposure," *The European Respiratory Journal*, vol. 11, pp. 1021-1027, Feb. 1998.

[38] W. F. Sensakovic, S. G. Armato 3rd, C. Straus, R. Y. Roberts, P. Caligiuri, A. Starkey, and H. L. Kindler, "Computerized segmentation and measurement of malignant pleural mesothelioma," *Medical Physics*, vol. 38, pp. 238-244, Jan. 2011.

[39] M. Rudrapatna, V. Mai, A. Sowmya, and P. Wilson, "Knowledge-Driven Automated Detection of Pleural Plaques and Thickening in High Resolution CT of the Lung," *Information Processing in Medical Imaging*, vol. 19, pp. 270-285, Jul. 2005.

[40] K. Chaisaowong, and T. Kraus, "Detection, modeling and matching of pleural thickenings from CT data towards an early diagnosis of malignant pleural mesothelioma," *In Proceedings of SPIE Medical Imaging*, vol. 9035, pp. 90350I, Mar. 2014.

[41] S. Singh, M. K. Kalra, M. D. Gilman, J. Hsieh, H. H. Pien, S. R. Digumarthy, and J. A. Shepard, "Adaptive Statistical Iterative Reconstruction Technique for Radiation Dose Reduction in Chest CT: A Pilot Study," *Radiology*, vol. 259, pp. 565-573, May 2011.

[42] Y. Ohno, D. TaKenaka, T. Kanda, T. Yoshikawa, S. Matsumoto, N. Sugihara, and K. Sugimura, "Adaptive Iterative Dose Reduction Using 3D Processing for Reduced- and Low-Dose Pulmonary CT: Comparison With Standard-Dose CT for Image Noise Reduction and Radiological Findings," *American Journal of Roentgenology*, vol. 199, pp. W477-W485, Oct. 2012.

[43] Y. Yamada, M. Jinzaki, T. Hosokawa, Y. Tanami, H. Sugiura, T. Abe, and S. Kuribayashi, "Dose reduction in chest CT: Comparison of the adaptive iterative dose reduction 3D, adaptive iterative dose reduction, and filtered back projection reconstruction techniques," *European Radiology*, vol. 81, pp. 4185-4195, Dec. 2012.

[44] The American College of Radiology, "ACR Practice Parameter for Performing and Interpreting Diagnostic Computed Tomography (CT)," https://www.acr.org/~/media/ADECC9E11A904B4D8F7E0F0BCF800124.pdf, amended 2014. Accessed February 21, 2017.

[45] J. Lee, and A. P. Reeves, "Segmentation of the airway tree from chest CT using local volume of interest," *In Proceedings of the Second International Workshop on Pulmonary Image Analysis (EXACT09)*, pp. 333-340, Sep. 2009.

[46] W. J. Kostis, A. P. Reeves, D. F. Yankelevitz, and C. I. Henschke, "Three-Dimensional Segmentation and Growth-Rate Estimation of Small Pulmonary Nodules in Helical CT Images," *IEEE Transactions on Medical Imaging*, vol. 22, pp. 1259-1274, Oct. 2003.

[47] J. Padgett, A. M. Biancardi, C. I. Henschke, D. F. Yankelevitz, and A. P. Reeves, "Local noise estimation in low-dose chest CT images," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, pp. 221-229, Mar. 2014.

[48] J. Lee, and A. P. Reeves, "Segmentation of Individual Ribs from Low-dose Chest CT," *In Proceedings of SPIE Medical Imaging*, vol. 7624, pp. 76243J, Feb. 2010.

[49] S. Liu, Y. Xie, and A. P. Reeves, "Segmentation of the sternum from low-dose chest CT images," *In Proceedings of SPIE Medical Imaging*, vol. 9414, pp. 941403, Mar. 2015.

[50] S. Liu, Y. Xie, and A. P. Reeves, "Automated 3D closed surface segmentation: application to vertebral body segmentation in CT images," *International Journal of Computer Assisted Radiology and Surgery*, vol. 11, pp. 789-801, 2016.

[51] A. P. Reeves, Y. Xie, and S. Liu, "Large-scale image region documentation for fully automated image biomarker algorithm development and evaluation," *Journal of Medical Imaging*, vol. 4, pp. 024505, 2017.

[52] Y. Xie, J. Padgett, A. M. Biancardi, and A. P. Reeves, "Automated aorta segmentation in low-dose chest CT images," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, pp. 211-219, Mar. 2014.

[53] Wapcaplet (E. Pierce), "Diagram of the human heart," http://commons.wikimedia.org/wiki/File:Diagram_of_the_human_heart_%28cropped%29.svg. Accessed June 27, 2016.

[54] P. D. Edwards, R. K. Bull, and R. Coulden, "CT measurement of main pulmonary artery diameter," *The British Journal of Radiology*, vol. 71, pp. 1018-1020, Oct. 1998.

[55] A. Devaraj, A. U. Wells, M. G. Meister, T. J. Corte, S. J. Wort, and D. M. Hansell, "Detection of pulmonary hypertension with multidetector CT and echocardiography alone and in combination," *Radiology*, vol. 254, pp. 609-616, Feb. 2010.

[56] Y. Xie, M. Liang, D. F. Yankelevitz, C. I. Henschke, and A. P. Reeves, "Automated measurement of pulmonary artery in low-dose non-contrast chest CT images," *In Proceedings of*

SPIE Med Imaging, vol. 9414, pp. 94141G, Mar. 2015.

[57] Y. Xie, M. D. Cham, C. Henschke, D. Yankelevitz, and A. P. Reeves, "Automated coronary artery calcification detection on low-dose chest CT images," *In Proceedings of SPIE Medical Imaging*, vol. 9035, pp. 90350F, Mar. 2014.

[58] BruceBlaus, "Illustration of coronary arteries," https://en.wikipedia.org/wiki/Coronary_circulation#/media/File:Blausen_0256_CoronaryArteries_02.png. Accessed June 20, 2017.

[59] R. Tota-Maharaj, P. H. Joshi, M. J. Budoff, S. Whelton, I. Zeb, J. Rumberger, M. Al-Mallah, R. S. Blumenthal, K. Nasir, and M. J. Blaha, "Usefulness of Regional Distribution of Coronary Artery Calcium to Improve the Prediction of All-Cause Mortality," *American Journal of Cardiology*, vol. 115, pp. 1229-1234, May 2015.

[60] X. Ding, P. J. Slomka, M. Diaz-Zamudio, G. Germano, D. S. Berman, D. Terzopoulos, and D. Dey, "Automated Coronary Artery Calcium Scoring from Non-contrast CT Using a Patient-specific Algorithm," *In Proceedings of SPIE Medical Imaging*, vol. 9413, pp. 94132U, Mar. 2015.

[61] Y. Xie, S. Liu, A. Miller, J. A. Miller, S. Markowitz, A. Akhund, and A. P. Reeves, "Coronary artery calcification identification and labeling in low-dose chest CT images," *In Proceedings of SPIE Med Imaging*, vol. 10134, pp. 101340L, Mar. 2017.

[62] S. A. Schwartz, M. S. Taljanovic, S. Smyth, M. J. O'Brien, and L. F. Rogers, "CT Findings of Rupture, Impending Rupture, and Contained Rupture of Abdominal Aortic Aneurysms," *American Journal of Roentgenology*, vol. 188, pp. W57-W62, Jan. 2007.

[63] S. Karazincir, A. Balci, E. Seyfeli, S. Akoglu, C. Babayigit, F. Akgul, F. Yalcin, and E. Egilmez, "CT assessment of main pulmonary artery diameter," *Diagnostic and Interventional Radiology*, vol. 14, pp. 72-74, Jun. 2008.

[64] Q. A. Truong, J. M. Massaro, I. S. Rogers, A. A. Mahabadi, M. F. Kriegel, C. S. Fox, C. J. O'Donnell, and U. Hoffmann, "Reference values for normal pulmonary artery dimensions by noncontrast cardiac computed tomography: the Framingham Heart Study," *Circulation: Cardiovascular Imaging*, vol. 5, pp. 147-154, Jan. 2012.

[65] A. Eisen, A. Tenenbaum, N. Koren-Morag, D. Tanne, J. Shemesh, M. Imazio, E. Z. Fisman, M. Motro, E. Schwammenthal, and Y. Adler, "Calcification of the thoracic aorta as detected by spiral computed tomography among stable angina pectoris patients: association with

cardiovascular events and death," *Circulation*, vol. 118, pp. 1328-1334, Sep. 2008.

[66] A. S. Agatston, W. R. Janowitz, F. J. Hildner, N. R. Zusmer, M. Viamonte Jr, R. Detrano, "Quantification of coronary artery calcium using ultrafast computed tomography," *Journal of the American College of Cardiology*, vol. 15, pp. 827-832, Mar. 1990.

[67] Y. Xie, Y, M. Htwe, J. Padgett, C. Henschke, D. Yankelevitz, and A. P. Reeves, "Automated aortic calcification detection in low-dose chest CT images," *In Proceedings of SPIE Medical Imaging*, vol. 9035, pp. 90350P, Mar. 2014.

[68] G. A. Rosito, J. M. Massaro, U. Hoffmann, F. L. Ruberg, A. A. Mahabadi, R. S. Vasan, C. J. O'Donnell, and C. S. Fox, "Pericardial fat, visceral abdominal fat, cardiovascular disease risk factors, and vascular calcification in a community-based sample: the Framingham Heart Study," *Circulation*, vol. 177, pp. 605-613, Feb. 2008.

[69] Y. Xie, M. Liang, D. F. Yankelevitz, C. I. Henschke, and A. P. Reeves, "Automated segmentation of cardiac visceral fat in low-dose non-contrast chest CT images," *In Proceedings of SPIE Medical Imaging*, vol. 9414, pp. 94140G, Mar. 2015.

[70] Vision and Image Analysis Group, "ELCAP Public Lung Image Database," http://www.via.cornell.edu/databases/lungdb.html. Accessed June 29, 2016.

[71] S. G. Armato 3rd, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. R. van Beek, D. Yankelevitz, A. M. Biancardi, P. H. Bland, M. S. Brown, R. M. Engelmann, G. E. Laderach, D. Max, R. C. Pais, D. P. Qing, R. Y. Roberts, A. R. Smith, A. Starkey, P. Batra, P. Caligiuri, A. Farooqi, G. W. Gladish, C. M. Jude, R. F. Munden, I. Petkovska, L. E. Quint, L. H. Schwartz, B. Sundaram, L. E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. V. Casteele, S. Gupte, M. Sallam, M. D. Heath, M. H. Kuhn, E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, B. Y. Croft, and L. P. Clarke, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans," *Medical Physics*, vol. 38, pp. 915-931, Feb. 2011.

[72] J. M. Bland, and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet*, vol. 327, pp. 307-310, Feb. 1986.

[73] International Early Lung Cancer Action Program, http://www.ielcap.org/. Accessed April 16, 2017.

[74] A. P. Reeves, Y. Xie, and A. Jirapatnakul, "Automated pulmonary nodule CT image characterization in lung cancer screening," *International Journal of Computer Assisted Radiology and Surgery*, vol. 11, pp. 73-88, Jan. 2016.

[75] The National Lung Screening Trial Research Team, "Reduced Lung-cancer Mortality with Low-Dose Computed Tomographic Screening," *The New England Journal of Medicine*, vol. 365, pp. 395-409, Aug. 2011.

[76] A. P. Reeves, A. B. Chan, D. F. Yankelevitz, C. I. Henschke, B. Kressler, and W. J. Kostis, "On measuring the change in size of pulmonary nodules," *IEEE Transactions on Medical Imaging*, vol. 25, pp. 435-450, Apr. 2006.

[77] A. C. Jirapatnakul, A. P. Reeves, T. V. Apanasovich, M. D. Cham, D. F. Yankelevitz, and C. I. Henschke, "Characterization of solid pulmonary nodules using three-dimensional features," *In Proceedings of SPIE Medical Imaging*, vol. 6514, pp. 65143E, Mar. 2007.

[78] R. J. Prokop, and A. P. Reeves, "A survey of moment-based techniques for unoccluded object representation and recognition," *CVGIP: Graphical Models and Image Processing*, vol. 54, pp. 438-460, Sep. 1992.

[79] A. C. Jirapatnakul, "Computer methods for pulmonary nodule characterization from CT images," *Master's dissertation,* Cornell University, Jan. 2011.

[80] S. A. Dudani, "The distance-weighted k-nearest-neighbor rule," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 6, pp. 325-327, Apr. 1976.

[81] T. Joachims, "Making large-scale support vector machine learning practical." *Advances in kernel methods*, MIT Press Cambridge, MA, pp. 169-184, 1999.

[82] T. Fawcett, "An introduction to ROC analysis." *Pattern Recognition Letters*, vol. 27, pp. 861-874, Jun. 2006.

[83] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, vol. 44, pp. 837-845, Sep. 1988.

[84] M. Salvatore, C. I. Henschke, R. Yip, A. Jacobi, C. Eber, M. Padilla, A. Knoll, and D. F. Yankelevitz, "Evidence of Interstitial Lung Disease on Low-Dose Chest CT Images: Prevalence, Patterns, and Progression," *American Journal of Roentgenology*, vol. 206, pp. 487-494, Mar.

2016.

[85] Y. Xie, M. Salvatore, S. Liu, A. Jirapatnakul, D. F. Yankelevitz, C. I. Henschke, and A. P. Reeves, "Identification of early-stage Usual Interstitial Pneumonia from low-dose chest CT scans using fractional high-density lung distribution," *In Proceedings of SPIE Medical Imaging*, vol. 10134, pp. 1013408, Mar. 2017.

[86] A. Depeursinge, A. S. Chin, A. N. Leung, D. Terrone, M. Bristow, G. Rosen, and D. L. Rubin, "Automated Classification of Usual Interstitial Pneumonia Using Regional Volumetric Texture Analysis in High-Resolution Computed Tomography," *Investigative Radiology*, vol. 50, pp. 261-267, Apr. 2015.

[87] D. J. Lederer, P. L. Enright, S. M. Kawut, E. A. Hoffman, G. Hunninghake, E. J. R. van Beek, J. H. M. Austin, R. Jiang, G. S. Lovasi, and R. G. Barr, "Cigarette Smoking Is Associated with Subclinical Parenchymal Lung Disease," *American Journal of Respiratory and Critical Care Medicine*, vol. 180, pp. 407-414, Sep. 2009.

[88] M. A. Gavrielides, L. M. Kinnard, K. J. Myers, J. Peregoy, W. F. Pritchard, R. Zeng, J. Esparza, J. Karanian, and N. Petrick, "Data From Phantom FDA. The Cancer Imaging Archive," http://doi.org/10.7937/K9/TCIA.2015.ORBJKMUX. Accessed April 16, 2017.

[89] M. A. Gavrielides, L. M. Kinnard, K. J. Myers, J. Peregoy, W. F. Pritchard, R. Zeng, J. Esparza, J. Karanian, and N. Petrick, "A resource for the assessment of lung nodule size estimation methods: database of thoracic CT scans of an anthropomorphic phantom," *Optics Express*, vol. 18, pp. 15244-15255, 2010.

[90] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository," *Journal of Digital Imaging*, vol. 26, pp. 1045-1057, Dec. 2013.